A Study and Comparative Analysis of Different Stemmer and Character Recognition Algorithms for Indian Gujarati Script

Rajnish M. Rakholia
PhD Scholar, R K University
Bhavnagar Highway
Rajkot – Gujarat, India

Jatinderkumar R. Saini Ph.D
Associate Professor and Director I/C
Narmada College of Computer Application
Bharuch – Gujarat, India

ABSTRACT
A lot of work has been reported on optical character recognition for various non-Indian scripts like Chinese, English and Japanese and Indian scripts like Tamil, Hindi Telugu, etc., in this paper, we present a literature review on stemmer, optical character recognition (OCR) and Text mining work on Indian scripts, mainly on the Gujarati languages. We have discussed the different techniques for OCR and text mining in Gujarati scripts, and summarized most of the published work on this topic and gives future directions of research in the field of Indian script.

General Terms
Stemmer, Gujarati character recognition

Keywords
Classification, feature extraction, Gujarati script, Gujarati stemmer, Indian script, pre-processing and segmentation.

1. INTRODUCTION
The Gujarati is an Indo-Aryan language and it is part of the Indo-European language family; it was adapted from the Devanagari script. Gujarati alphabet mainly includes 34 consonants (ornamented sounds) and 14 vowels (pure sounds). Gujarati language does not having a horizontal bar for its letterforms but shares some appearances as that of Devanagari, Sanskrit, Marathi etc.

Gujarati language is very popular language and more than 50 million people speak Gujarati language. Gujarati is the official language of the Gujarat state of India, however many national organizations such as Banks, use English and Gujarati. Even all the documents in the government offices of Gujarat state usually appear in these two languages. Gujarati uses the virama to form conjunct characters which mean "lame" in Gujarati.

Gujarati language and script developed in three distinct phases - 10th to 15th century, 15th to 17th century and 17th to 19th century. The first phase is marked by use of Prakrit, Apabramsa. In second phase, 'Old Gujarati' script was in wide use; the script first appeared in print in a 1797 advertisement. The third phase is the use of script developed for ease and fast writing [31].

1.1 Framework of Gujarati Symbol
Character in Gujarati language can be logically divided into the following parts based on the position of the shapes involved.

1.1.1 Baseline Area: this portion contains consonants and independent vowels.

1.1.2 Area below and above the Baseline: used for below-base and above-base dependent vowels respectively.

1.1.3 Area before and after the Baseline: this is the *placeholder* for consonants and independent vowels [31].

2. STEMMER FOR GUJARATI SCRIPT
Stemming is the process to transform the words in texts into their grammatical root form.

Sheth J and Patel B (2014) suggested DHIYA a stemmer for Gujarati language, EMILLE corpus is used for training and evaluation of the stemmer's performance. They obtained accuracy of 92.41% [26].

In ***(Sheth and Patel, 2012)*** they discussed different stemming techniques for Gujarati language. It was found that POS tagged data and hand-crafted rules are improved the performance of English stemmer. Thus they suggest that building a POS tagger for Gujarati language will also improve Gujarati stemming and apart of POS and hand-crafted rules, the researchers have also used unsupervised technique.

They suggest that currently the work done on stemming for Gujarati language is measured in terms of correctness but not in terms of incorrectness i.e. overstemming and understemming. They proposed a model for Gujarati Stemmer which not just focuses on traditional approach but also incorporates naïve approaches to derive an accurate and efficient stemmer. In which they make GUJ Stem-suffix frequency table (GSFT) to store the frequency of stem-suffix pair based on the training done on a Gujarati Corpus. Gujarati Language Rule Set (GLRS) store the rules for suffix stripping. These rules are defined based on the study of the characteristics of Gujarati language. They proposed a GUJ_STRIP_function which is based on artificial neural network technique along with the probability and smoothing concept [27].

Chauhan K, Patel R and Joshi H (2013) stemming is well known information retrieval technique which is used for text retrieval system. They used Gujarati stemmer in information retrieval of Gujarati text. They applied a rule-based approach as per Gujarati Morphology and select data set from Gujarati news paper and identified various possible word suffixes in corpus. In their methodology they select a class of words that share common suffix of given character and replace each of them by their original root.

In their experiments, to evaluate the retrieval performance, the mean average precision (MAP) values were considered. They evaluated the results separately for title (T), combination of title and description (TD) and the combination of title,

description and narration (TDN). The investigations of their experiments show that using of stemmer in Gujarati text documents contribute to a significant amount of increase in precision values in information retrieval tasks and improve the performance nearly by 13% [11].

Ameta J, Joshi N and Mathur I (2011) were proposed rule base stemmer algorithm for Gujarati language. They created a test corpus of 3000 Gujarati words and listed all possible suffix for each word in advanced. They applied this algorithm on test data, if suffix of executing word is match with given suffix list then it will be remove and root word will be store. In these 3000 words; they had 389 words which were unique. Among these 389 words 218 stems had more than one morphological variants present in the corpus and 171 stems had only one morphological variant. The stemmer is able to capture most of the morphological variants. They tested their systems for the verification and claim that; they got good accuracy for same [2].

Patel P, Popat K and Bhattacharya P (2010) they presented hybrid approach for lightweight stemmer for Gujarati language. Instead of using unsupervised approach they made hand-crafted Gujarati suffix list in order to improve performance of stemmer. They used the EMILLE corpus for training and evaluating the stemmer's performance. After experiment they extracted around 10 million words from corpus and these words also contained Gujarati transliterations of English words.

They tried to filter out these words by using a Gujarati to English transliteration engine and an English dictionary. Afterwards they divided extracted data in to five equal partitions of which four are for training and one for testing. They used five-fold cross validation for evaluating the performance. The results of this experiment clearly indicate that there is a large improvement in the performance of the stemmer with the use of hand-crafted suffixes and minimum stem size. The use of hand-crafted suffixes boosted the accuracy of their stemmer by about 17% and helped us achieve an accuracy of 67.86 % [25].

Ameta J, Joshi N and Mathur I (2013) were used Part-of-Speech Tagging and stemmer assisted transliteration for Gujarati-Hindi machine translation. Transliteration means mapping of source language text into the target language. Simple mapping decreases the efficiency and accuracy of translation system. They could not get effective result in direct transliteration without any rules or constraints, the main reason behind of that is suffix is get attached with root word and it decrease the efficiency of system. They created a raw corpus of 5400 POS-tagged sentences and used 202 stemming and tagging rules to assist transliteration. The POS-Tagged corpus is a collection of text files. They tested their system on a total of 5000 sentences. Result shown that, 54.48% of Gujarati words translation and transliteration are same. The efficiency of our transliteration scheme is about 90% [1].

Patel M and Balani P (2013) clustering is preprocessing for stemming; they proposed clustering algorithm for Gujarati testing, segmented characters are taken from scanned document by pre-processing stage. The Hopfield Neural Network with 900 input neurons and 900 output neurons are used for classification. Principal Component Analysis (PCA) is used for features extraction and MATLAB tool is used for this purpose. Result shown that overall recognition rate is 93.25% and the error rate 6.751% [29].

Kamal et al (2013) proposed feed forward back propagation neural network for the classification of the language. They used 5000 tagged words set and try to cluster those Gujarati word base on proposed algorithm; and they found this algorithm works properly with tagging words and got accuracy 98%. But this algorithm is not giving proper result with untagged words; if words is not properly tagged then there are prey good chances of error in clustering process otherwise on correctly tagged words they found accuracy more than 98%. For the Gujarati language they used WX notation and all implementations are done by using JAVA language [24].

3. FEATURES EXTRACTION FOR GUJARATI SCRIPT

Features extraction is the process to extract information that is related to particular object or group of objects, base on this information we can classify in particular category.

Baheti M and Kale K (2012) used affine invariant moments approach to extracting features for offline isolated handwritten Gujarati numerals. To classify the Gujarati numerals, they used Principal Component Analysis (PCA), Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Gaussian distribution function for features extraction. As compared to overall recognition accuracy, SVM has shown recognition accuracy of 92.28%, Gaussian distribution function shown 87.2%, and K-NN classifier has shown 90.04%, while PCA shown 84.1%. After analyzing of study SVM proves to be better classifier than PCA, K-NN and Gaussian distribution function classifier for affine invariant moments as feature extraction technique [5].

Thakar H and Kumbharana C (2014) they present analysis of structural features and its classification for consonants of Gujarati language. This study is related to hand written character or printed consonants in Gujarati language. According to them each consonant has unique structural feature which distinguishes it from other consonants. Structural features that are font independent, size independent, works well even with shape distortion. Eight Structural features are selected as a base for categorizing consonants into twenty two groups from F1, F2,…..F22 [30].

Patel C and Desai A (2011) they presented to extract the words from lines in Gujarati script to improve the accuracy of OCR system for Gujarati language. They used some combined methods like projection profile with morphological operations to enhance accuracy of the word extraction [23].

4. CLASSIFICATION AND RECOGNITION OF GUJARATI CHARACTERS AND NUMERALS

This is next phase of features extraction to identity an object and classify in different categories.

Solanki P and Bhatt M (2013) applied Hopfiled Neural Network in Optical Character Recognition (OCR) for printed Gujarati script. They took 748 images for training data set in which two images per character and font size 12 point. For

Gujarati numerals. Before classification, binarization and skeletonization are done in preprocessing stage of handwritten numerals. They trained this network for total 30 sets of digits (number of digits are 300) and tested for other 60 sets of digits (number of digits are 600) they obtained success rate 100% and 80.33% respectively for Gujarati handwritten digit identification. They applied the binarization on the character, to eliminate the various intensities of gray pixels of the image to make binary .

Choksi A and Thakkar S (2012) used Fuzzy-kNN algorithm for recognition of similar appearing Gujarati characters in scanned documents. Many Gujarati character has similar shape and characteristic that is affected on accuracy of character recognition, to handle this problem two different features Geometric and Wavelet are used in pair. Train data set is prepared by typing Gujarati characters in different font types and size and then scanned, while testing data (in which similar characters are appear) collected from different sources like Gujarati news paper, book and scanned document . Result of Wavelet features (Fuzzy k-NN) are almost 100%, while accuracy obtained for similar appearing characters with k-NN were 67% and General Regression Neural Network were near about 97%. Less recognition efficiency were observed in printed and scanned Gujarati newspapers and documents because of broken characters and insufficient print quality[12]

Chaudhari S and Gulati R (2011) proposed model for OCR to read and separate digits which is written in English and Gujarati script. This proposed system is only used for numeric character for Gujarati and English language. The digits in Gujarati and English are based on sharp curves and hardly any straight line is available, means each digit will identified by specific characteristic, shape and curve.

The model is divided in three basic image processing stages like pre-processing, segmentation, and feature extraction and classification. Reading image file, Binarization, and noise removal will be done in Pre-Processing stage. After Pre-processing, the segmentation is performed in which, first the line segmentation is performed and then using connected component analysis character segmentation is performed. In last stage of this process identify the features of each character and classify in to Gujarati and English digits. They obtained 98.30% for Gujarati Digits and 98.88% for English digits at character level [10].

Mamta M and Kale K (2011) proposed Support Vector Machine (SVM) to recognition of Gujarati numerals and used affine invariant moments for feature extraction. Each isolated numeral is segmented into blocks for computing the features. They created sample data set manually by taking handwriting imprints and then scanned with resolution of 300dpi. Through the morphological techniques they removed noise or skew to enhance the image in pre-processing phase. By using any unbiased algorithm image can be normalize in same size, they used nearest neighbor interpolation technique to normalized image in size of 40x40. They dealt with feature-based recognition of handwritten characters by using morphological dilation and skeletonization. Success rate for this approach were 90.55% [20].

Patel C and Desai A (2011) they presented process of identification of various zone for hand written Gujarati character based on distance transform. The zone identification is used to extract modifiers from root character, mainly they focused on three zones middle, upper and lower zone. In which middle zone contain root (basic) character while upper zone and lower zone store modifier of Gujarati character and it will change meaning of word(s) in Gujarati language. Zone identification is important phase of Optical Character Recognition (OCR) to improve accuracy of OCR system.

Desai A (2010) Used feed forward back propagation neural network to classify Gujarati numerals and they created five patterns for each digit in both clockwise and anticlockwise directions. They took 278 sets of various digits in which 11 sets were created by a standard font. They recorded the success rate for standard fonts 71.82%, for handwritten training sets as 91.0% while for testing sets 81.5% success rate was recorded [22].

Dholakiya et al (2005) presented an algorithm to identify various zones for Gujarati printed text; in that algorithm they considered horizontal and vertical profiles for each character [13]. In *(Dholakia et al., 2007)* used wavelet features, GRNN classifier and KNN classifier for the printed Gujarati text, they achieved success rate of 97.59% and 96.71% respectively [14].

5. ANALYSIS OF RELATED WORK ON INDIAN SCRIPT

Table-1 gives Summary and comparison of various classifiers, features extraction technique and accuracy of related work for Indian script, mainly for Gujarati script.

Table-1

Authors	Input Script	Pre-processing	Feature extraction	Classifier	Accuracy
Solanki and bhatt (2013)	Printed Gujarati script	character segmentation	PCA is used for features extraction	Hopfiled Neural Network	93.25%
Kamal et al (2013)	Gujarati numerals	binarization and skeletonization	-	feed forward back propagation neural network	80.33%
Baheti and kale (2012)	Handwritten Gujarati numerals	Binarization, bounding box segmentation and skeletonization	Affine invariant moments	SVM	92.28%
				K-NN	90.04%
				PCA	84.10%
Choksi and thakkar (2012)	Gujarati scanned documents	-	Geometric and Wavelet features	Fuzzy-kNN	100 %
Bag et al (2011)	Bangla Handwritten characters	Image binarization, Character-level segmentation;	Structural convexity	LCS technique	60.25%
Chaudhari and Gulati	Numeric character for Gujarati and	Binarization, and noise	hybrid statistical and	template matching	98.30%(Guj)

(2011)	English language	removal	structural feature	classifier	98.88%(Eng)
Mamta and Kale (2011)	Gujarati numerals	morphological dilation and skeletonization	affine invariant moments	Support Vector Machine	90.55%
Goswami et al (2011)	Printed Gujarati characters	SOM as Pre-Processing	not require prior feature identification stage	Self Organizing Map (SOM) with k-NN	97.50%(Training)
					82.36%(Testing)
Desai A (2010)	Handwritten Gujarati numerals	Created five patterns for each digit in both clockwise and anticlockwise directions.	-	feed forward back propagation neural network	81.5%
Singh et al (2009)	Old Hindi (Devanagari) Handwritten characters	Skeletonization, Normalization and compression	Gradient features	Neural network	95.00%
Kompalli et al (2009)	Devanagari Printed words	Graph-based character segmentation	Gradient and GSC features	Neural network, *K*-nearest neighbour, and SFSA classifiers	95.50%
Majumdar (2007)	Bangla Printed characters	Thinning; Thickening	Curvelet coefficient features	K-nearest neighbor classifier	96.80%
Banashree & Vasanta (2007)	Devanagari Handwritten numerals	Image binarization; Directed graph construction	Features from end-point information	Neuro-memetic model classifier	92.00%
Dholakia et al (2007)	Printed Gujarati text	-	wavelet features	GRNN	97.59%
				KNN	96.71%
Bhattacharya et al (2006)	Bangla Handwritten characters	Smoothing; Binarization; Removal of extra long headlines	Chain code histogram features	MLP classifier	92.14 %
Bhattacharya et al (2002)	Bangla Handwritten numerals	Noise removal; Vector skeletonization	Topological and statistical features	Hierarchical tree and MLP classifiers	93.26%

We conclude from this analysis, neural network is most frequently used classifier for Indian script mainly in Gujarati, Bangla and Devnagari (old Hindi). We acquired average accuracy for Gujarati script as 90.58%, while for Bangla script 85.61% and we obtained 94.70% for Devnagari script. We also analyzed in this study, that binarization and skaletionization are used for pre-processing stage in most of research.

6. CONCLUSION

This paper studies various types of stemmer and other pre-processing step for Gujarati language as well as various methodology, features extraction techniques and classifiers to recognize characters and digits written in Indian language. Due to complex structure of Gujarati framework and complexity of Gujarati grammar, still it is scope in research to improve accuracy of stemmer for Gujarati language and recognize hand written character in Gujarati script.

7. REFERENCES

[1] Ameta J, Joshi N and Mathur I, " Improving the quality of Gujarati-Hindi machine translation thought Part-of-Speech tagging and stemmer assisted transliteration", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3, 2013.

[2] Ameta J, Joshi N and Mathur I, "A Lightweight Stemmer for Gujarati", Department of Computer Science, Apaji Institute, Banasthali University, Rajasthan, India.

[3] Antani S, Agnihotri L, "Gujarati Character Recognition", Proceeding 5th ICDAR, IEEE Computer Society, 1999, pp. 418-422.

[4] Bag S, Bhowmick P and Harit G, "Recognition of Bengali handwritten characters using skeletal convexity and dynamic programming, Proceeding of International Conference on Emerging Applications of Information Technology, 2011, pp. 265–268.

[5] Baheti M and Kale K, "Gujarati Numeral Recognition: Affine Invariant Moments Approach", Special Issue of International Journal of electronics, Communication & Soft Computing Science & Engineering, 2012.

[6] Banashree N and Vasanta R, "OCR for script identification of Hindi (Devnagari) numerals using feature sub selection by means of end-point with neuro-memetic model". Int. J. Intell. Tech. 2:2007, pp. 206–210.

[7] Bhattacharya et al, "A hybrid scheme for handprinted numeral recognition based on a self-organizing network and MLP classifiers". Int. J. Pattern Recognition Artificial Intelligence.16: 2002, pp.845–864.

[8] Bhattacharya et al, "Recognition of handprinted Bangla numerals using neural network models", Proceedings of the AFSS International Conference on Fuzzy Systems, 2002, pp. 228–235.

[9] Bhattacharya U, Shridhar M and Parui S, "On recognition of handwritten Bangla characters", Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing,2006, pp. 817–828.

[10] Chaudhari S and Gulati R, "Character Level Separation and Identification of English and Gujarati Digits from Bilingual (English-Gujarati) Printed Documents", Proceedings published in International Journal of Computer Applications (IJCA), 2011.

[11] Chauhan K, Patel R and Joshi H, "Towards Improvement in Gujarati Text Information Retrieval by Using Effective Gujarati Stemmer", Journal of Information, Knowledge and Research in Computer Engineering, Vol. 2, Issue 2, 2013.

[12] Choksi A and Thakkar S, "Recognition of Similar appearing Gujarati Characters using Fuzzy-KNN Algorithm", International Journal of Computer Applications, Volume 55– No.6, 2012.

[13] Dholakia J, Negi A and Rama M, "Zone Identification in the Printed Gujarati Text", Proceeding of 8th ICDAR IEEE Computer Society, 2005, pp. 272-276.

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

[15] Dholakia J, Yajnik A and Negi A, "Wavelet Feature Based Confusion Character Sets for Gujarati Script" proceeding of ICCIMA, IEEE, 2007, pp. 366-379.

[16] Dobariya A and Rathod V, "Comparative Study of Different Classifier for Gujarati off-Line Text Recognition", International Journal for Scientific Research & Development, Vol. 2, Issue 01, 2014.

[17] Goswami M, Prajapati H and Dabhi V, "Classification of Printed Gujarati Characters using SOM based K-Nearest Neighbor Classifier", Pattern Recognition, Image Processing & Computer Vision, Proceeding of ICIIP, IEEE, 2011, pp. 1-5.

[18] Kompalli S, Setlur S and Govindaraju V, "Devanagari OCR using a recognition driven segmentation framework and stochastic language models", Int. J. Doc. Anal. Recognit. 12: 2009, pp. 123–1308.

[19] Majumdar A, "Bangla basic character recognition using digital curvelet transform", Journal of Pattern Recognition Research 2:2007, pp. 17–26.

[20] Mamta M and Kale K, "Support Vector Machine based Gujarati Numeral Recognition", International Journal on Computer Science and Engineering (IJCSE), Volume 3 - No.7, 2011.

[21] Moro et al, "Gujarati Handwritten Numeral Optical Character through Neural Network and Skeletonization", Jurnal Sistem Komputer, Indonesia, Volume 3 - No.1, 2013.

[22] Patel C and Desai A, "Segmentation of text lines into words for Gujarati handwritten text", Proceeding of ICSIP, IEEE, 2010, pp. 130-134.

[23] Patel C and Desai A, "Zone Identification for Gujarati Handwritten Word", Proceeding of 2nd EAIT, IEEE, 2011, pp. 194-197.

[24] Patel M and Balani P, "Clustering Algorithm for Gujarati Language", International Journal for Scientific Research & Development (IJSRD) Vol. 1, Issue 3, 2013.

[25] Patel P, Popat K and Bhattacharyya P, "Hybrid Stemmer for Gujarati", 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010.

[26] Sheth J and Patel B, "Dhiya: A stemmer for morphological level analysis of Gujarati language", Proceeding of ICICT, IEEE, 2014, pp. 151-154.

[27] Sheth J and Patel B, "Stemming Techniques and Naïve Approach for Gujarati Stemmer", International Journal of Computer Applications (IJCA), 2012.

[28] Singh D, Dutta M and Singh S, "Neural network based handwritten Hindi character recognition system", Proceedings of the Bangalore Annual Compute Conference, article no. 15, 2009.

[29] Solanki P and Bhatt M, "Printed Gujarati Script OCR using Hopfield Neural Network", International Journal of Computer Applications, Volume 69 - No.13, 2013.

[30] Thaker H and Kumbharana C, "Analysis of Structural Features and Classification of Gujarati Consonant for Offline Character Recognition", International Journal of Scientific and Research Publications, Volume 4, Issue 8, August 2014.

[31] Gujarati language origin: http://en.wikipedia.org/wiki/Gujarati_alphabet

8. AUTHOR'S PROFILE

Rajnish Rakholia received his Master's degree in Computer Applications from Ganpat University in 2010 and Bachelor of Science in Physics from Saurashtra University in 2007. Currently he is working as an Assistant Professor in MCA department of S. S. Agrawal Institute of Computer Science, Navsari and pursuing Ph.D. in Computer Science from R K University, Rajkot. His areas of interest are Web Technologies, Text Mining, Sen timent Analysis and Opinion Mining.

Dr. Jatinderkumar R. Saini is Ph.D. from VNSGU, Surat. He secured First Rank in all three years of MCA and has been awarded Gold Medals for this. Besides being University Topper, he is IBM Certified Database

Associate (DB2) as well as IBM Certified Associate Developer (RAD). Associated with more than 50 countries, he has been the Member of Program Committee for more than 50 International Conferences (including those by IEEE) and Editorial Board Member or Reviewer for more than 30 International Journals (including many those with Thomson Reuters Impact Factor). He has more than 55 research paper publications and nearly 20 presentations in reputed International and National Conferences and Journals. He is member of ISTE, IETE, ISG and CSI. Currently he is working as Associate Professor and Director I/C at Narmada College of Computer Application, Bharuch, Gujarat, India. He is also Director (Information Technology) at Gujarat Technological University, Ahmedabad (GTU)'s A-B Innovation Sankul.