Original Article

Journal of Clinical Forensic Sciences
Volume 2 Issue 2, July – December 2024

Offline Script Identification from Handwritten Gujrati Script Documents

Akash Sharma[1], Chhote Raja Patle[2], Anuwanshi Sharma[3], Anita Yadav[4]

How to cite this article:
Akash Sharma, Chhote Raja Patle, Anuwanshi Sharma *et al.* Offline Script Identification from Handwritten Gujrati Script Documents. Jr of Clin Forensic Sci. 2024;2(2):59–64.

Abstract

This study focuses on offline script identification of handwritten Gujarati script documents using Optical Character Recognition (OCR) techniques. The goal is to develop an efficient system capable of accurately identifying the Gujarati script from handwritten documents. The process begins with the collection of a diverse dataset of offline handwritten Gujarati script documents. The dataset includes various handwriting styles to ensure the model's adaptability. Ground truth labels are annotated for training and evaluation purposes.

Preprocessing techniques are employed to enhance the image quality of the handwritten documents. These techniques involve noise removal, image resizing, and normalization, resulting in clearer and standardized input for the subsequent steps. OCR techniques are then applied to perform the script identification task. These techniques involve the extraction of features and patterns specific to the Gujarati script from the pre-processed images. Machine learning algorithms, such as Support Vector Machines (SVM) or Convolutional Neural Networks (CNN), are trained on the extracted features to learn the script identification patterns. The trained model is evaluated using standard performance metrics, including accuracy, precision, recall, and F1 score. The dataset is divided into training and testing sets to assess the model's effectiveness in identifying the Gujarati script. Once the model is trained and evaluated, it can be deployed for practical use. Given an input handwritten document, the OCR system utilizes its learned patterns to accurately identify and classify the Gujarati script. Overall, this study presents a concise approach to offline script identification of handwritten Gujarati script documents using OCR techniques. The proposed system shows promise in accurately reorganizing the Gujarati script, paving the way for further advancements in this field.

Keywords: Offline Script Identification; Handwritten; Gujarati Script; Document; OCR; Optical Character Recognition; Dataset; Preprocessing; Feature Extraction; Machine Learning; Support Vector Machines; SVM; Convolutional Neural Networks; CNN; Performance Evaluation; Accuracy; Precision; Recall; F1 Score.

Author's Affiliation: [1]M.Sc Student, [2]Assistant Professor, [4]Associate Professor, Department of Forensic Science, Sanjeev Agrawal Global Educational University, Bhopal 462022, Madhya Pradesh, India, [3]PhD Scholar, Department of Forensic Science, Galgotias University, Greater Noida 203201, Uttar Pradesh, India.

Correspondence: Anita Yadav, Associate Professor, Deparment of Forensic Science, Sanjeev Agrawal Global Educational University, Bhopal 462022, Madhya Pradesh, India.

E-mail: anitakakas7@gmail.com

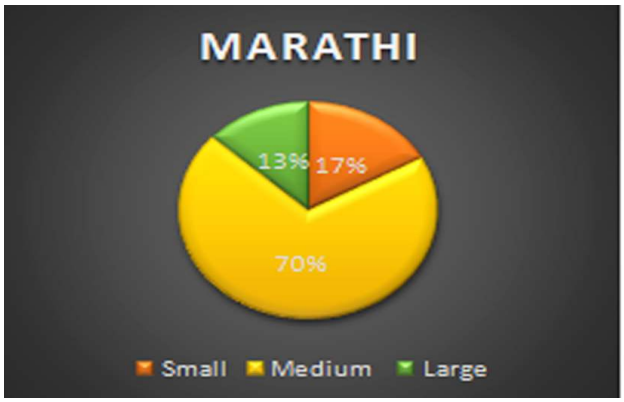
Received on: 09.08.2023 **Accepted on:** 01.11.2023

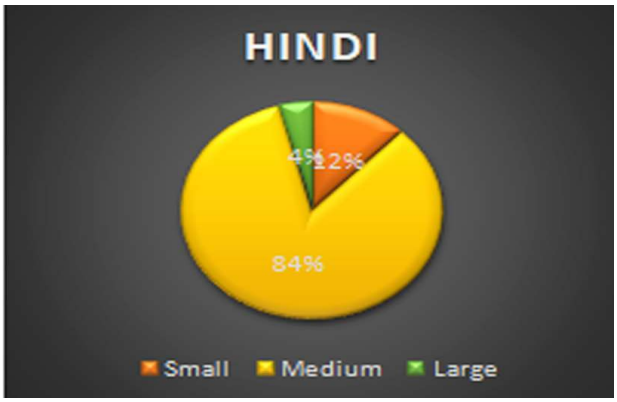
INTRODUCTION

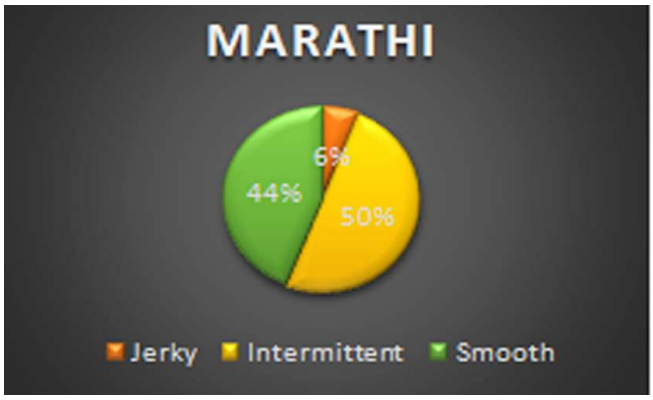
Handwriting is an ancient form of communication that has evolved over time. It allows people to document various aspects of life, including history, culture, literature, law, science, and mathematics. India has ten main scripts for official communication, including Devanagari, Bengali, Gurumukhi, Gujarati, Oriya, Kannada, Telugu, Tamil, Malayalam, and Urdu (Nastaliq). Most Indian scripts descended from

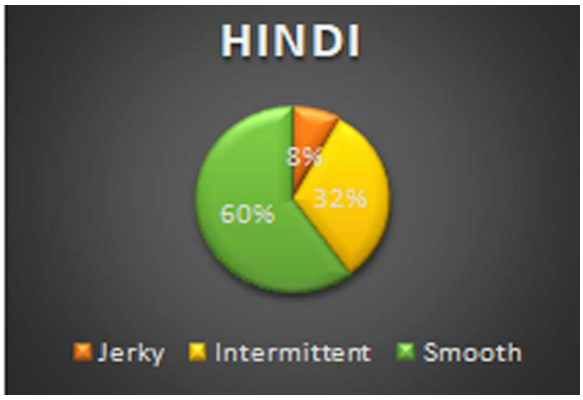
the ancient Brahmi script, with Devanagari being the official language. Devanagari is not considered a regional script.

Regional scripts in India predominantly contain compound characters, numbers, vowels, and consonants. These characters are created by combining two or more consonants, with complex shapes. Over 22.02% of postal documents in West Bengal are written in Bangla script. The policy of state administrations in India encourages official transactions in regional languages, and some office memos and instructions are handwritten in regional scripts. Gujarati, the phonetic language of Western India, is recognized by union territories and Gujarat, with around 50 million speakers. The fundamental character set of Gujarati script consists of 36 consonants and 13 vowels called Swar and vyanjan.









Gujarati script, a type of Indian script, uses over 250 consonant clusters and vowel modifiers to represent vowels. It shares similarities with Devanagari letters and does not discriminate between lower and upper case. Gujarati also lacks the Shirorekha (headline) and differs from Devanagari in terms of consonants and vowels. A Dependent Vowel Modifier (Matra) is added to each consonant or conjunct to indicate the presence of a dependent vowel. These Matras can appear before, after, above, or below the core consonant or conjunct, and can take on different forms depending on the related consonant. Handwritten scripts are widely used in daily life, and automatic computer recognition is needed to identify specific words or character sequences. A system for automatic recognition of handwritten Gujarati scripts is crucial in today's rapidly advancing world. This technology aims to convert manuscripts and other handwritten materials into electronic, machine-readable documents, reducing the need for human effort. Offline Handwritten Character Recognition (HCR) is a technique that uses computers to recognize handwritten, printed, or type writer - produced text. Offline HCR applications include automatic reading for postal mail sorting, bank checks, administrative offices, scene text identification, and electronic preservation of handwritten records. The importance of data in today's data-driven world is undeniable, and transforming old manuscripts and other handwritten documents into machine-editable formats is essential.

OCR phases include pre-processing, which involves converting a document into images or files using cameras or scanners, and thresholding. This process converts the image into a binary image

using global and adaptive thresholding methods. Skew detection is used to recognize skewed images' angles, and other steps may be performed based on research needs. Segmentation is a versatile approach used by most OCRs, dividing the picture into lines, words, and characters for easier recognition.

Line segmentation is a process that divides a picture or document into lines using appropriate algorithms. Word segmentation divides the image into individual words, while character segmentation removes each character from the segmented word. Various methods and methodologies exist for different levels of segmentation. Recognition is the most crucial phase, using various approaches and strategies to recognize characters in the script. Feature extraction is a procedure that processes each character using techniques like Template Matching, Zoning, and Transformations.

Classification and recognition are followed by analyzing the extracted features using methods like Nearest Neighbor (NN), Euclidean Distance, Neural Network, and Support Vector Machine (SVM). The final stage involves obtaining classified data in text format or a document file, with post processing steps like error detection, rectification, and grouping for better results.

Currently, neural networks and template matching are commonly used in handwritten character recognition. However, template matching is ineffective due to font discrepancies, font slants, stroke connections, and stroke breaks. A popular NN method based on feed forward neural networks and back propagation learning is developed for OCR problems. The training phase involves a potential input and the expected output of the network. The information content within the borders is digitalized according to the Neural Network algorithm's requirements using an 8 × 8 grid.

The process of character recognition involves feature extraction, classification, and recognition using various methodologies and classifiers. Common approaches include Nearest Neighbor (NN), Euclidean Distance, Neural Network, and Support Vector Machine (SVM). Text data/ document is obtained, and post processing steps like error detection, rectification, and grouping are performed for better outcomes. Currently, neural networks and template matching are commonly used algorithms in handwritten character recognition. However, template matching is not effective in certain situations, such as font discrepancies, font slants, font defilements, stroke connections, and stroke breaks.

A popular NN method for OCR is developed based on feed forward neural networks with back propagation learning. The training phase involves two components: a potential input and the expected network's output. The network can be given any input and produce an output to determine the type of pattern presented to the network. The information content within the borders is digitalized according to the Neural Network algorithm's requirements by creating an 8 × 8 grid over the area. A cross correlation function is used to find patterns similar between input test images and trained database images. The window size is typically fixed at 8 x 8, but can be changed if necessary.

The demand for human machine interaction has increased the potential of character recognition systems for real-world applications. Chinese, Latin, Arabic, and Japanese scripts have received the most attention in this field, while work on Indic scripts is still in its infancy.

Challenges in Gujarati OCR

Gujarati OCR faces challenges due to its complicated script, including similar characters, font styles, conjuncts, skewed characters, and broken personalities. These issues impair accuracy and quality in output.

Handwritten character recognition (HCR) has been studied in optical character recognition and pattern recognition. In India, a new approach, pattern matching, is proposed for character identification in Gujarat's official language, Gujrati. This technique uses preprocessing and picture enhancement techniques to identify different handwritten characters, with neural networks enhancing precision.[1] Developed an offline handwritten Gujarati numeral database using stroke features from 14,000 samples from 140 individuals. The database recognizes Gujarati and Devanagari numerals using a novel technique extracting low-level stroke features. The database was tested using SVM classifiers and k-NN classifiers.[2] Offline Script Identification (OSI) simplifies applications like automatic document archiving, online and offline document browsing, and script-specific Optical Character Recognition (OCR) in multilingual settings. Singh PK. conducted a study on OSI approaches for Indian scripts in 2015. This survey discusses feature extraction and categorization methods related to OSI in Indic scripts. It will benefit policymakers and practitioners, connecting researchers studying

various Indic scripts, and provide a foundation for future research operations.[3] Office and government are increasingly paperless, leading to the development of OCR (Optical Character Recognition) to convert paper documents into digital text. HCR is a type of OCR designed for reading handwritten text, while PCR focuses on printed text. OCR can be online or offline, with both formats being recognized offline or online. This article aimed to create an offline HCR system for Gujarati using artificial intelligence. The study collected 10,000 photos from 250 people, using CNN and MLP for character recognition. The main goal was to develop a continuous workflow for word-level image to text conversion.[4] In the digital age, digitizing paper information requires converting it to ASCII or Unicode. This studied offline handwritten Gujarati character recognition using structural elements and a decision tree classifier. The study involved digitization, preprocessing, structural features extraction, and recognition, achieving a success rate of 88.78% for five Gujarati script characters.[5]

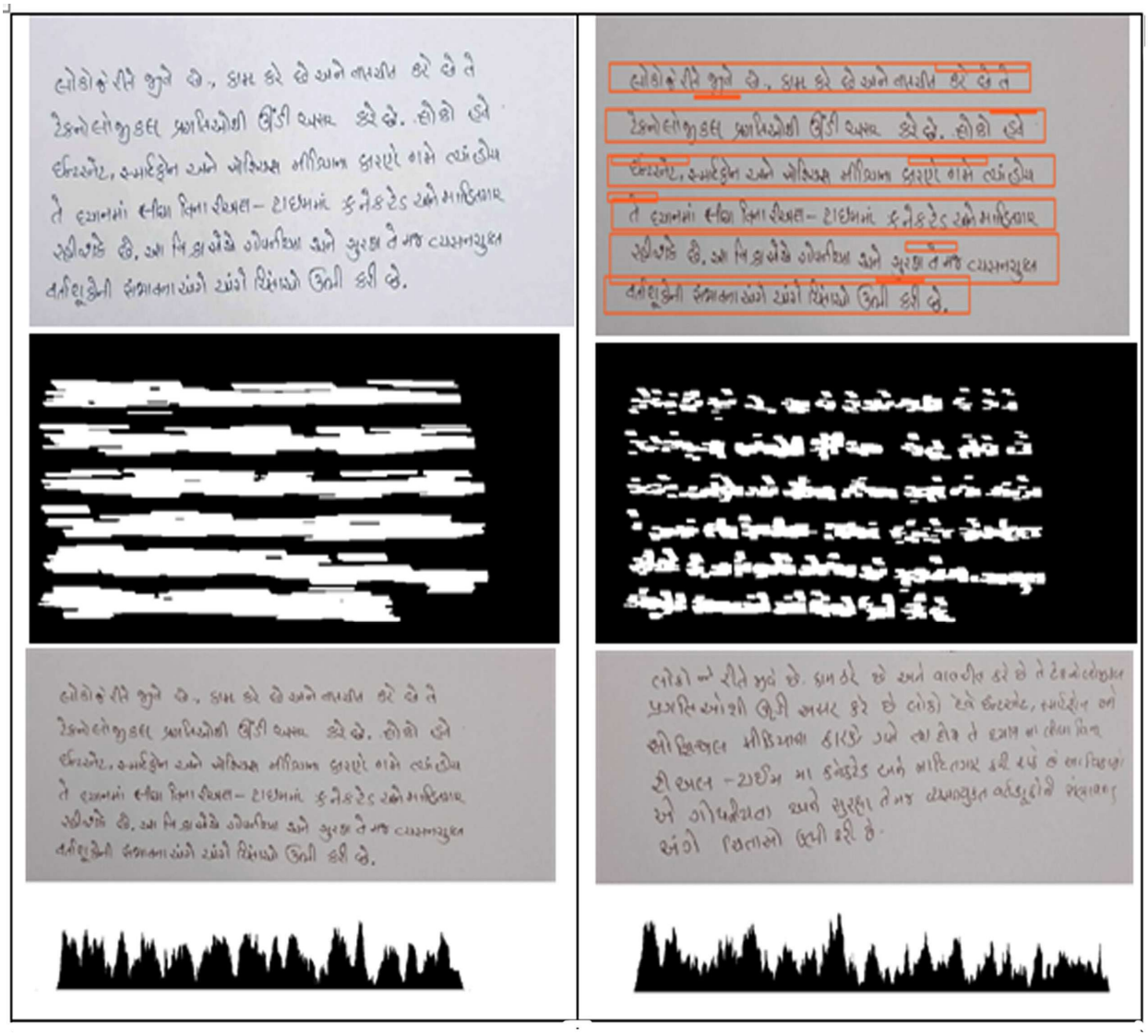


Fig. 1: Sample Processing using OCR

Prasad *et al.* proposed a preprocessing method for recognizing Gujarati characters using a median filter, thinning characters, and template matching. The method achieved an average recognition rate of 71.66%.[6]

METHODOLOGY

The study focuses on development of an efficient system for offline script identification

in handwritten Gujarati script documents using advanced techniques and algorithms. The study also analyze the impact of technological advancements on social interactions, economy, education, and healthcare. It examines opportunities and challenges, offering insights on harnessing benefits while mitigating risks. The study investigates emerging technologies like AI, automation, and digital platforms, aiming to inform policymakers, practitioners, and the public about responsible technology implementation.

The samples were collected samples from diverse settings (home, office, college, etc.) to accommodate a range of writing styles. The total number of people involved in this data collection activity was a hundred. The main distinguishing feature of our developed database is its heterogeneity in three critical factors: data collection locations, the educational attainment of writers who participated in data collection, and the native language of writers who participated in data collection.

Finally, a dataset of one hundred handwritten document samples is constructed, with a maximum of one hundred document pages per script. Each document page in the script databases is defined by the height, breadth, aspect ratio, total number of text lines and words, and statistical assessments of the normal horizontal and vertical stroke widths. The handwritten document pages are collected from a variety of persons who have been requested to write the material or just translate the passage into their native language.

RESULTS & DISCUSSIONS

Table 1: Characteristics of handwriting after processing through OCR

Sample Parameters	Height (n=100)	Width (n=100)	Number of text line (n=100)	Word Segment (n=100)	Line Segment (n=100)	Character Segment (n=100)	Aspect Ratio (n=100)	Precision (n=100)
Mean	9.47	17.21	7.57	29.5	7.75	355.73	1.971	98.71
Mode	8	18	8	19	8	357	2.25	99
Median	9	17	8	25.5	8	357	1.8	99
Range (Min-Max)	4-14	12-20	4-11	7-78	5-12	345-360	1.14-5	98-99

After processing the images through OCR the study showed mean height as 9.47 with median 9 and mode 8 with a range of 10. Similarly mean width obtained was 17.21 with 18 & 17 are mode and medians respectively.

In the same mean values of number of text line is 7.57 followed by mean values of word segment and line segment was 29.5 and 7.75 respectively. The mean value of character segment was 355.73 with aspect ratio value as 1.971 and with a precision of 98.71.

The study describes the creation of benchmark datasets for unconstrained handwritten document pages that contain Guajarati script words. Total hundred handwritten document pages in Gujarati script had been collected. Each document contains various people characters, text, numerals, and other symbols. In addition, evaluating a word segmentation technique on mixed-script document pages written in Gujarati Script. The study will help to achieve the goal of bringing together researchers working on various Indian scripts. It is considering that the recent improvements in Indian regional scripts also provide a better foundation for future research endeavours.

CONCLUSION

This study examines handwritten Indic script recognition for optical character recognition in multilingual and multi-script environments. Various script features have been proposed for identification, with page level data achieving the best recognition accuracy. Visual appearance based features outperform structural based features at all levels of script recognition, resulting in significantly improved categorization scores. The study highlights the applicability of the suggested classifier combination approach to handwritten Indic script recognition.

REFERENCES

1. Prasad JR, Kulkarni UV, Prasad RS. Offline handwritten character recognition of Gujrati script using pattern matching. In 2009 3rd international conference on anti-counterfeiting, security, and identification in communication 2009 Aug 20 (pp. 611-615). IEEE.

2. Goswami MM, Mitra SK. Offline handwritten Gujarati numeral recognition using low-level strokes. International Journal of Applied Pattern Recognition. 2015;2(4):353-79.
3. https://arxiv.org/ftp/arxiv/papers/2009/2009.07435.pdf accessed on 20/07.2023
4. https://www.researchgate.net/profile/Vinayak-Vinay/publication/360946905_Language_Translation_for_Impaired_People_using_NLP_Semantics/links/6294e4d2c660ab61f852a211/Language-Translation-for-Impaired-People-using-NLP-Semantics.pdf accessed on 23/07/2023.
5. Savani M, Vadera D, Limbachiya K, Sharma A. Character Segmentation from Offline Handwritten Gujarati Script Documents. In Information and Communication Technology for Competitive Strategies (ICTCS 2021) ICT: Applications and Social Interfaces 2022 Jun 23 (pp. 61-70). Singapore: Springer Nature Singapore.
6. Pareek J, Singhania D, Kumari RR, Purohit S. Gujarati handwritten character recognition from text images. Procedia Computer Science. 2020 Jan 1;171:514-23.