ABSTRACT

In the current digital era, due to very heavy usage of computers, smartphones and computing devices in all aspect of human life, it is expected that information/document has to be in e-form. So, such devices should recognize native languages to help common people to perform their daily tasks. Also, the paper is a very comfortable and feasible medium to store data, there is a great demand for the software techniques that can automatically extract, analyse and store information from the physical handwritten documents for later retrieval. But communication with the computer using this natural means of paper is not possible due to high variation of handwriting and large character set for the Indian languages like Gujarati. Handwritten text recognition for the Indian scripts is still in the need of maturity for some of the languages like Gujarati.

Recognizing handwritten character of Indian scripts are more challenging due to its large character set and nature of character shapes. In the presented work we have formulated offline handwritten text recognition system capable of recognizing isolated characters and text for Indian languages. The aim is to investigate the Hand-written text recognition for Gujarati script that can be generalized for Indian scripts. In this research, we have model the architectural solution which consisting of preprocessing, segmentation and recognition phases. Standard handwritten text document dataset is not available for the experiment. We have built dataset consist of 107 documents.

Preprocessing of image document is essential for reduction in data and for improving quality for better processing. Also, information in text document is best represented as foreground and background. We have identified preprocessing phase task as converting image to grayscale, removing noise, binarization, cropping and skew detection and correction. After the preprocessing text document image is transform into binary image which is then segmented in next phase.

Segmentation is the most challenging phase for text recognition system. In scripts like Gujarati, have diacritic marks corresponding to vowels and used in combination of consonants makes segmentation problem more challenging. Hence

the selection of appropriate segmentation strategy is very important step in construction of text recognition system. One of the major objectives of this work is to study different approaches used in text segmentation and address issues in segmentation of Gujarati handwritten text document.

Segmentation approach used in this research is based on combination approach. It uses partial projection histogram for line detection combined with RLSA to strengthen the histogram. Knowledge of constructing Gujarati Akshara is used for line tracing using parameters obtained from document after applying connected component analysis. The output of segmentation classified to core characters, and diacritic marks. The performance of line segmentation for Gujarati text document is 98.32%, character segmentation is 82.39% tested on 107 documents comprise of 1135 lines. We used structural characteristics and position for recognition of diacritic marks using tree classifier.

Isolated Gujarati handwritten character recognition module is implemented to recognise core characters using feed forward neural network. The dataset consisting of 13500 sample images of isolated Gujarati characters is used to recognize 54 character class consists of 49 basic Gujarati alphabets, numerals, and 5 character parts. In preprocessing of IGHCR, we have limited preprocessing task based on the noise in dataset. The new features namely partial radial histogram and water reservoir area methods are used in this research for character recognition. We have modelled different features vector to measure effectiveness of features. The overall highest recognition rate obtained is 91.49% for 54 classes.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATION

ABBREVIATION

Abbreviation	Full form
ACH	Average Character Height
ACW	Average Character Width
ANFC	Adaptive Neuro-Fuzzy Classifier
ANN	Artificial Neural Network
CC	Connected Component
CEDAR	Centre of Excellence for Document Analysis and Recognition
CENPARMI	The Centre for Pattern Recognition and Machine Intelligence
CLAHE	Contrast Limited Adaptive Histogram Equalization
DCT	Discrete Cosine Transform
DeitY	Department of Electronics And Information Technology
DPI	Dots Per Inch
EL	Endpoints and its Location
HCR	Handwritten Character Recognition
HTR	Handwritten Text Recognition
HWR	Height to Width Ratio
ICDAR	International Conference in Document Analysis and Recognition
IGHCR	Isolated Gujarati Character Recognition
JPEG	Joint Photographic Experts Group
k-NN	K-Nearest Neighbours
LH	Length of Horizontal Line
LV	Length of Vertical Line
MLP	Multilayer Perceptron
MNIST	Modified National Institute of Standards and Technology
NB	Number of Branch Points

NIST	National Institute of Standards And Technology
NL	Number of Loops
NLP	Natural Language Processing
OCR	Optical Character Recognition
OGHTR	Offline Gujarati Handwritten Text Recognition
PCA	Principal Component Analysis
PCR	Printed Character Recognition
PNG	Portable Network Graphics
PRH	Partial Radial Histogram
RGB	Red Green Blue
RLSA	Run Length Smoothing Algorithm
SE	Structuring Element
SVM	Support Vector Machine
SW	Stroke Width
WRA	Water Reservoir Area