IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-6, Issue-11, Nov 2018 E-ISSN: 2347-2693

A Review On Offline Gujarati Word Categories Using Hybrid Features

[1*]Amitkumar T Solanki, [2]Sheshang D Degadwala, [3]Kishori Shekokar

[1, 2, 3]Computer Engineering Department, Sigma Institute of Engineering, Vadodara, Gujarat, India

Available online at: www.ijcseonline.org

Accepted: 17/Nov/2018, Published: 30/Nov/2018

***Abstract*—** The input to handheld devices using a traditional keyboard is not a user-friendly process for Indian scripts due to large and complex character sets. Handwritten character recognition can be a best possible solution. Handwritten character recognition is gaining noteworthy attention in the area of pattern matching and machine learning. Named categories Recognition is a method to find for a particular Named field from a file or an image, recognize it and classify it into specified Entity Classes like Name, Location, Organization, Numbers and Other Categories. The main purpose of using Hybrid Feature is that it provides better performance and can be easily implemented for any languages. A remarkable amount of work has been carried out for many languages like English, Greek, and Chinese etc. But still a wide scope is open for Indian Origin Languages like Hindi, Gujarati, and Devanagari etc. As Gujarati is not only the Indian Language, but a language that is most spoken in Gujarat. Thus, in this research paper different type of features and classification techniques are compare with advantages and disadvantages. After review all methods give idea for future direction research for Guajarati OCR recognition.

***Keywords*—** Gujarati Language, Named Entity Recognition, Invariant feature Extraction, Character classification.

I. INTRODUCTION

Offline handwritten word recognition is a widely studied pattern recognition problem and has direct applications in automated check processing, handwritten postal mail sorting, automatic processing of handwritten forms etc. The problem can be solved using three possible approaches, namely incremental, holistic and hybrid [1]. In the incremental approach the word image is further divided into segments and uses incremental model for recognition whereas in the holistic approach the entire word is considered as a single unit of recognition. The hybrid approach is a mix of two. In this paper, holistic approach for word recognition has been proposed to identify handwritten city names in Gujarati script. Gujarati is 26th most-spoken native language in world with 65.5 million Gujarati speakers all over the world. The language also has a rich collection of literary work including the handwritten notes by M. K. Gandhi [2]. There are 13 vowels and 34 consonants. Apart from vowels and consonants, Gujarati word construction also contains diacritics (Matras) because of which, character recognition in Gujarati language becomes difficult. This challenge can be addressed by recognizing the whole word. However, such system can only be used in the domain specific problems where the size of the vocabulary is limited. For example, automated postal address sorting using city names or automated processing of handwritten forms etc. The implementation of word recognition system includes pre-processing of word images, segmentation of words, feature extraction, and classification of word images.

Implementation of word recognition technique on the Gujarati database makes the work easy, as it does not require extraction of symbols and glyphs from the word image. In this paper, we describe a handwritten Gujarati word recognition technique using Histogram of Oriented Gradients (HoG) features and state of the art classifier like Support Vector Machine (SVM) and k-Nearest Neighbor (kNN). The experiments were performed on a moderate sized database of handwritten Gujarati city names. The work produced has direct application in handwritten postal address processing.

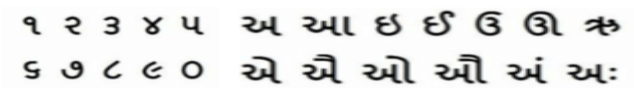

Figure 1. Gujrati Characters

Section I contains the introduction of basic approach for weather forecasting. II contain the related works of basic literature papers. Section III contain the methodology and algorithms section IV explain the comparative study between different algorithms and at last conclusion and future scope.

II. RELATED WORK

Parita R. Paneri, Ronit Narang, Mukesh M.Goswami propose technique for Offline Handwritten Gujarati Word Recognition Gujarati language is an Indo-Aryan language that has a complex structure wherein extracting each character becomes hectic because of the presence of diacritics. Implementation of word recognition technique on the Gujarati database makes the work easy as it does not

require extraction of symbols and glyphs from the word image. In this paper, we describe a handwritten Gujarati word recognition technique using Histogram of Oriented Gradients (HoG) features and state of the art classifier like Support Vector Machine (SVM) and k- Nearest Neighbor (kNN). The experiments were performed on a moderate sized database of handwritten Gujarati city names. The work produced has direct application in handwritten postal address processing. Texture features are not using only hog features are used so misclassification cannot be handled [1].

Vishal A. Naik, Apurva A. Desai propose technique for Online Handwritten Gujarati Character Recognition Using SVM, MLP, and And K-NN. This paper tried to present a we present a system to recognize online handwritten character for the Gujarati language. Support Vector Machine (SVM) with linear, polynomial & RBF kernel, k-Nearest Neighbor (k-NN) with different values of k and multi-layer perceptron (MLP) are used to classify strokes using hybrid feature set they have trained using 3000 samples and achieved a maximum accuracy of 91.63% with SVM and minimum accuracy of 86.72% with MLP. Provides low accuracy for high resemblance, similar characters and confusing characters. Misclassification for confusing characters [2].

Komil Vora, Dr. Avani Vasant, Rachit Adhvaryu propose work for Named Entity Recognition and Classification for Gujarati Language. This paper presenting Named Entity Recognition (NER) is a method to search for a particular Named Entity (NE)[1] from a file or an image, recognize it and classify it into specified Entity Classes like Name, Location, Organization, Numbers and Others Categories. It is the most useful element of the technique known as Natural Language Processing (NLP) which makes text extraction very easy [2]. In this paper, we focus on using Hidden Markov Model (HMM) based techniques to recognize the Named Entity (NE) for Gujarati language. The main aim of using HMM is that it provides better performance and can be easily implemented for any languages. A remarkable amount of work has been carried out for many languages like English, Greek, and Chinese etc. But, still a wide scope is open for Indian Origin Languages like Hindi, Gujarati, and Devanagari etc. As Gujarati is not only the Indian Language, but a language that is most spoken in Gujarat. Thus, in this paper, we emphasis on proposing a NER based scheme for Gujarati Language using HMM but Optimum output is not provided in all cases. Only character is recognized [3].

Chhaya C Gohell, Mukesh M Goswam, Yishal K Prajapate propose technique for On-line Handwritten Gujarati Character Recognition Using Low Level Stroke In this paper, we described an online HCR system for Gujarati characters based on stroke based features for the dataset of 4500 samples. The combination of LLS features and directional features was shown which gives the good accuracy for online HCR of Guajarati characters and numerals but it have some limitations like Low accuracy, Big future vector and high processing time, Works for handwritten character only[4].

Apurva A. Desai work for Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space. This work presents an algorithm for handwritten Gujarati alphabet recognition. This work shows that for Gujarati handwritten alphabet identification hybrid feature set is more effective than simple structural feature set. Also SVM with polynomial kernel (c = 2) gives the best accuracy of Gujarati handwritten alphabet identification compare to other classifiers like kNN and SVM with Gaussian kernel. But with low identification accuracy [5].

S. S. Magare, Y. K. Gedam, D. S. Randhave, R. R. Deshmukh proosed work for Character Recognition of Gujarati and Devanagari Script. In this paper, we describe the different techniques of character recognition for Gujarati and Devanagari script. Character recognition is usually referred to as OCR. Review of this paper will provide a way for researcher to develop a tool for Gujarati and Devanagari script recognition. This paper describes basics of character recognition, its type, challenges associated with it and the special properties of Gujarati and Devanagari script but it have some limitation it takes high processing time, and low accuracy [6].

III. METHODOLOGY

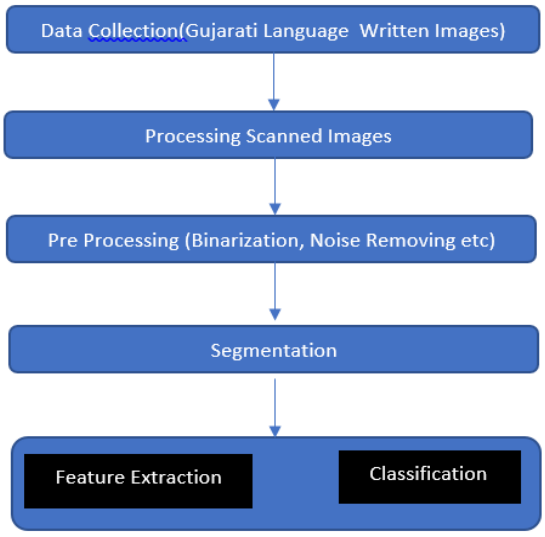

Figure 2. Basic Steps For object mining and tagging algorithm

A. Preprocessing :

During the time spent character picture acknowledgment, it requires some preprocess for acknowledgment making pictures more helpful for mechanized acknowledgment they are expecting to refine. This refinement is known as

preprocessing. There is diverse strategy for preprocess. Binarization is the way toward changing over grayscale picture in to double (Black and White) picture, with the goal that picture information will just contain 0 and 1. Binarization procedure is generally utilized for isolating closer view from foundation utilizing required dimension of thresholding. Computerized picture comprises of assortment of clamors. These commotions are required to be expelled from a picture for better handling. Morphological activity, Median channel and Weiner channel are utilized to expel commotion from a picture. Middle channel lessens obscuring of edges. Diminishing and Filling Smoothing infers both Filling and Thinning. Diminishing decreases width of character while Filling takes out hole, little breaks and gaps in digitized character. To acquire characters of uniform size, revolution and inclination Normalization is connected on picture. To enhance the precision of character acknowledgment Normalization lessens shape variety. Amid the digitization of report page, usually picture isn't adjusted accurately, or it might occur by human while composing archive. To make in effectively adjust Skew discovery and remedy method is utilized. Skew identification procedure can be arranged in to gatherings: Analysis of Projection profile, Hough change, grouping, associated part and connection between's line methods

B. Segmentation

Division of a picture is the way toward subdividing picture into number of parts. Division accepts the frame as Paragraph Segmentation, Line Segmentation, Word Segmentation and Character Segmentation. Passage astute division partitions the report into section. Line astute division separates section into line. Line astute division can utilize a flat projection profile-based systems Word shrewd division separates line into word. At last, Character savvy division isolates words into characters. Chain code histogram can be utilized for each fragment. Even projection document technique is utilized for division. [6]

C. Feature extraction

The feature extraction starts from an initial set of measured values and builds derived values intended to be informative, non-redundant, and should facilitating the subsequent learning. The histogram of oriented gradients (HOG) is a feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image and hence describes local shape of an object. The histogram is computed for each of the dense grid of uniformly spaced and non-overlapping cells. [1]

i. Histogram of oriented gradients [HOG]

The histogram of oriented gradients (HOG) is a feature descriptor used in computer vision and image processing for the purpose of object detection. The technique counts occurrences of gradient orientation in localized portions of an image. Although the printed Gujarati character recognition has been continually researched over the past few decades, it still requires the performance improvement for applying to real applications. One of the main problems is the diversity of new and old Gujarati character fonts, i.e., the newer fonts are created, the more recognition errors are increased. Furthermore, HOG counts occurrences of gradient orientation in localized portions of an image. The essential thought behind the HOG descriptors is that local object appearance and shape within an image can be described by the distribution of intensity gradients or edge directions. This technique divides the image into small square cells and then computes the histogram of gradient directions or edge directions based on the central differences. HOG features are calculated by taking orientation histograms of edge intensity in a local region. In this paper, HOG features are extracted from all locations of a grid on the digit image as candidates of the feature vectors. [1]

ii. Rule Based Approach

Rule based approaches are the most efficient for NER systems. It generally uses rules written manually by the experts. It provides high accuracy as compared to other approaches. Basically Rule based approach is classified into following:

a) Linguistic Approach

b) List Look-up Approach

The main drawback of Rule Based Approaches are: 1). A vast expertise is required for generating rules for a domain. 2). An implementation of a system requires effort and time very much. 3). The rules defined for one language or domain cannot be transferred or used for other language or domain. 4). Implementing a small modification is very complex. [3]

Machine Learning Based Approach

Machine Learning based approaches are fully dependent on arithmetical models to predict Named Entity in each document. A great amount of Metadata information is required to make this approach successful and worthy. [3]

iii. Hidden Markov Model

A hidden Markov model (HMM) is an algebraic Markov model where it is assumed that the system is a developed with overlooked states. A HMM can be presented as the simplest dynamic Bayesian network. A Hidden Markov Model (HMM) can be considered as an outline of a fusion model where the overlooked variables which controls the combination component that is taken for observation are through markov process rather independent. Hidden Markov models are mostly known for their applications such as speech & handwriting recognition, movement recognition, part of-speech labeling, musical score following, bioinformatics etc. HMM is proficient of allocating semantic labels to tokens over a inputs; this is beneficial for text related tasks that involve some ambiguity, including part-of-speech labeling, text separation, named entity recognition and data mining tasks. However, most of the natural language processing tasks are reliant on determining an entity associated with information. An example would be

word "Rajkot" is a Location; therefore word "Rajkot" will receive a tag of Location. [3]

D. Classification

In the last stage of character recognition is uses different methods like KNN, SVM and ANN to classify characters. K-Nearest Neighbor Classifier (k-NN) is a distance based classification method. It computes a Euclidean distance between testing data with all data samples of a training set. K samples having minimum distance with testing sample will be selected from the training set. The class label for a majority of k-nearest data samples from training set will be set for a testing sample. Different k values lead to different processing and results. A smaller value may lead to low accuracy and higher value may lead to more processing. [2]

SVM uses to classify handwritten Gujarati alphabets. It is a binary classifier which classifies a group into two classes. Support vector machine creates a hyperplane in n-dimensional place separates two classes. Support vector machine places hyperplane at a distance from where the nearest point of both the classes have the largest distance. This distance is known as ‘‘margin’’. Support vector machine gives efficient results if all parameters are set properly. Therefore, it is important to understand various parameters which play an important role in SVM. [5]

IV. COMPARATIVE STUDY

Table I. Comparison between Classification Extraction Method

Classifier	Advantage	Limitation
SVM[7]	SVM is less complex. Produce very accurate classifiers. Less over fitting, Robust to noise.	SVM is binary classifier, to do a multi-class classification, pair-wise classifications can be used Computationally expensive, thus runs slow
ANN[4]	Ann can perform tasks which linear program cannot. When element of neural network fails it continue to work.	They do not classify and cluster data, a lot of chips and a distributed run-time to train on very large datasets.
KNN[8]	-Robust to noisy training data -Effective if the training data is large	-Distance based learning is not clear which type of distance to use and which attribute to use to produce the best result. -computation cost is quit high

Table II. Comparison between Classification Extraction Method

Feature	Advantage	Limitation
HOG[7]	HOG (Histogram of Oriented Gradients) can be used to detect any kind of objects, as to a computer, an image is a bunch of pixels and you may extract features regardless of their contents.	Complete processing is time consuming.
Chain Code[6]	The advantage is reduction in storage volume	The disadvantage is loss of generality. The basic chain code is very sensitive to noise & it is not rotationally invariant.
HMM[2]	Strong statistical foundation Efficient learning algorithms-learning can take place directly from raw sequence data. Allow consistent treatment of insertion and deletion penalties in the form of locally learnable Can handle inputs of variable length-they are the most flexible generalization of sequence profiles.	HMMs often have a large number of unstructured parameters. First order HMMs are limited by their first-order markov property. They cannot express dependencies between hidden states. The HMM is unable to capture higher order correlation among features.

V. CONCLUSION AND FUTURE SCOPE

After review different types of features and classifiers we can conclude that future research presents the first ever attempt for OCR Gujarati word recognition. In that we can use invariant shape feature as well ML classifier. Also recognize City and Name categories for useful to other researchers in the field.

REFERENCES

[1] Parita R. Paneri, Ronit Narang, Mukesh M.Goswami, “Offline Handwritten Gujarati Word Recognition”, 2017 Fourth International Journal on Image Information Processing (ICIIP).

[2] Vishal A. Naik, Apurva A. Desai “ Online Handwritten Gujarati Character Recognition Using SVM, MLP, And K-NN ” , 8th ICCCN IIT Delhi,(2017).

[3] Komil Vora, Dr. Avani Vasant, Rachit Adhvaryu, “Named Entity Recognition And Classification For Gujarati Language” Intl.

Conference on Advances in Computing, Communications and Informatics (ICACCI)(2016).

[4] Chhaya C Gohell, Mukesh M Goswam, Yishal K Prajapate, "On-line Handwritten Gujarati Character Recognition Using Low Level Stroke" Third International Conference on Image Information Processing(2015).

[5] Apurva A. Desai "Support vector machine for identification of handwritten Gujarati alphabets using hybrid feature space" CSI Publications, (2015).

[6] S. S. Magare, Y. K. Gedam, D. S. Randhave, R. R. Deshmukh "Character Recognition of Gujarati and Devanagari Script" International Journal of Engineering Research & Technology (IJERT), (2014).

[7] Mukesh M. Goswami, Suman K. Mitra "High-Level Shape Representation in Printed Gujarati Characters" Sixth International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017).

[8] Swital J. Macwan, Archana N. Vyas "Classification of Offline Gujarati Handwritten Characters" IEEE(2015).

[9] Ami Mehta, Ashish Gor "Multi font Multi size Gujarati OCR with Style Identification" International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS-2017)

Authors Profile

Amitkumar T Solanki completed Bachelor of Engineering in Computer Science & Engineering from Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, Madhya Pradesh. He is pursing Master of Computer Engineering from Sigma institute Engineering affiliated by Gujarat Technological University Gandhinagar. He has published 2 research papers on Security issue and concurrency control of Distributed database Transaction System and Cloud Computing in International Journal for Scientific Research & Development. He has an experience in industry for 1 year as a Database developer and in 3.5 years' experience as lecturer.

Dr. Sheshang D. Degadwala Completed Ph.D. in Computer Engineering from Madhav University, Abu Road, Sirohi, Rajasthan, India in year 2018. He is currently working as Head of Computer Engineering Department in, Sigma Institute of Engineering, Vadodara, India since 2012. He has published more than 58 research papers in reputed international journals and 3 in National conferences including Thomson Reuters and conferences including IEEE, Springer and it's also available online. His main research work focuses on Image Processing, Information Security and Data Mining. He has 6 years of teaching experience and 6 years of Research Experience.

Kishori Shekokar completed Master of Engineering in Computer Science & Engineering form Amravati university, Maharastra in year 2009. She is currently working as Assistant Professor in Computer Department , Sigma Institute of Engineering Vadodara. She has published more than 19 research papers in reputed international journals. Her main research work focuses on Image Processing, Datamining, Cloud Computing. She has 10.5 years of teaching and research experience.

