Survey on Offline Character Recognition for Handwritten Gujarati Text

Bhumika B. Patel
Dept. of Information Technology
SVM Institute of Technology
Bharuch, India

Hinaxi M. Patel
Dept. of Computer Engineering
SVM Institute of Technology
Bharuch, India

ABSTRACT
Handwritten character recognition (HCR) is one of the most interesting topics in image processing and pattern recognition. Handwritten Character Recognition is not a difficult task for humans, but for a machine it's really tough. In Handwritten Character Recognition method the input is scanned from images, documents and real time devices like tablets, computers and digitizers etc. which are then interpreted into digital text. There are basically two approaches - Online Handwritten recognition which takes the input at run time and Offline Handwritten Recognition which works on scanned images. Offline handwritten character recognition is a process where the computer understands automatically the image of handwritten script. Several applications are used HCR like mail sorting, bank processing and document reading, etc. Offline handwriting recognition is comparatively difficult, as different people have different handwriting style. This research is confined within offline character recognition only. Gujarati is the name of script used to write the Gujarati language and spoken by the people in state Gujarat in India.

General Terms
Handwritten character recognition (HCR), Support Vector Machine (SVM).

Keywords
Handwritten character recognition (HCR), Optical Character Recognition (OCR), Support Vector Machine (SVM).

1. INTRODUCTION
The development of handwriting recognition systems began in the 1950's when there were human operators whose job was to convert data from various documents into electronic format, making the process quite long and often affected by errors [1]. Automatic text recognition aims at limiting these errors by using image preprocessing techniques that bring increased speed and precision to the entire recognition process. Handwriting recognition has been one of the most fascinating and challenging research areas in field of image processing and pattern recognition in the recent years. It contributes immensely to the advancement of automation process and improves the interface between man and machine in numerous applications. Optical character recognition is a field of study than can encompass many different solving techniques. Neural networks, support vector machines, deep learning and statistical classifiers seem to be the preferred solutions to the problem due to their proven accuracy in classifying new data [1].

The Optical Character Recognizer actually is on vector which translates handwritten text images to a machine based text. In general, handwriting recognition is classified into two types as off-line and on-line. In the off-line recognition, the writing is usually captured optically by a scanner and the completed writing is available as an image. Online Handwritten Text is written by a stylus on a tablet. This method used which laser, inkjet devices, for obtaining machine printed text [1].

There is extensive work in the field of handwriting recognition, and a number of reviews exist. These methods can be useful for real-time applications like OMR Sheet correction when it's filled in Gujarati character through OCR, Bank cheque processing, Birth Certificate, mail sorting, document reading and postal address recognition. Moreover, in the off-line systems, the neural networks and support vector machines have been successfully used to yield comparably high recognition accuracy levels. As a result, the off-line handwriting recognition continues to be an active area for research towards exploring the newer techniques that would improve recognition accuracy. Therefore, for this paper, we would decide to work on an off-line handwritten character recognition using system Support Vector Machine (SVM) [1].

Many researchers have presented their work in the area of character recognition for English and Arabic script. Observation based on preliminary literature review indicates some work for South Indian script also, whereas very few researches works is traced for character recognition in Gujarati script, which is an official language of Gujarat state, Western part of India. Paper is organized into different sections as previous work, Set of Gujarati consonants, Methodology for proposed work as structural feature selection, analysis in form of decision table for classification of Gujarati consonants.

Fig1: Main Components of Character Recognition

2. COMPONENT OF CHARACTER RECOGNITION
The basic mechanism of offline character recognition consists of following phases: Image Pre-processing, Feature Extraction, Classification and Post Processing. Figure 1 shows main components of Character Recognition.

In pre-processing scanned document is converted to binary image and various other techniques to remove noise, to make it ready and appropriate for feature extraction are applied. These techniques include segmentation to isolated individual characters, skeletonization, contour making, normalization, filtration etc [2].

In Feature extraction is the important step in character recognition, however other steps also need to be optimized

because these steps are closely related to each other as outputs of earlier step is inputted to later step. Feature extraction is used to extract the most relevant information which is used to classify the objects. Most relevant is in the sense to minimize the within class pattern variability and to maximize the between class variability. Features can be broadly classified into two categories: structural features and statistical features. Structural features are involved of structural elements like loop, line, crossing point, curve, end point and stroke etc. Statistical features are computed by some statistical operations on image pattern and these include features like zoning, projection, profiling, histogram and distance etc. Structural and statistical features appear complementary to each other and many other features can be derived from the basics of these features [2].

Decomposing the feature extraction phase in two sub-phases- feature construction and feature selection. In feature construction raw features and even irrelevant features are considered not to lose any information. Adding all these features increases the dimensionality of patterns. In feature selection step only relevant features are identified and selected to create feature vectors. Each pattern having feature vector is classified in predefined classes using classifiers. Classifiers are first trained by a training set of pattern samples to prepare a model which is later used to recognize the test samples (see figure 1). The training data should consist of wide varieties of samples to recognize all possible samples during testing. Some examples of generally practiced classifiers are- Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Artificial Neural Network (ANN) and Probabilistic Neural Network (PNN) [2].

In post processing step we bind up our work to create complete machine encoded document through the process of recognition, assigning Unicode values to characters and placing them in appropriate context to make characters, words, sentences, paragraphs and finally whole document [2].

3. RELATED WORK

Tivedi et al. [3] proposed that Systems classified according to data acquisition techniques (Online & Offline) and Systems classified according to the text type (Printed & Handwritten). Its recognize Gujarati Characters from Scanned Image. And Applied different classifiers namely SVM, k-NN and Naïve-Bayes classifier for Recognition of constants. Performance of difference classifiers measured using 10-fold cross validation.

Macwan et al. [4] proposed that Different algorithms From different domains have been considered for comparative analysis like Transform Domain (DWT, DCT and DFT), from Spatial Domain; Geometric Method (Gradient feature), Structural method (Freeman chain code) and Statistical method (Zernike Moments).Recognizing Gujarati Characters from Scanned Handwritten numeric character Image. Transform domain gives good accuracy especially DFT but it takes more time. Structural and Statistical method does not give good accuracy individually. While a combination of structural and statistical method provides higher accuracy in less time. In geometric method, gradient features give higher accuracy in less time.

Mehula et al. [5] proposed that First, the need for segmentation is justified in the context of text based information retrieval. Then, the various factors affecting the segmentation process are discussed. Followed by the levels of text segmentation are explored. Finally, the available techniques with their superiorities and weaknesses are reviewed, along with directions for quick referral are suggested. To segment a text based image. The pixel counting approach is the increased computation and the resulting space complexity, thereby experiencing diminution in computational speed.

Prasad et al. [6] proposed that moment based normalization process for character images for the purpose of enhancing the performance of character recognize for isolated characters of Gujarati. Moment based image normalization applied in the pre-prepossessing stage of a character recognition task. Apply a normalization procedure to the image so that it meets a set of predefined moment criteria. A recognition rate of 86.6 % for the isolated characters of Gujarati is achieved.

Sharma et al. [7] proposed that Recognition of isolated Gujarati handwritten characters is proposed using three different kinds of features and their fusion. Chain code based, zone based and projection profiles based features are utilized as individual features. Show substantial enhancement over state-of-the-art and authenticate our proposal. Increase in number of classes can make the problem difficult and fusion of some more significant features may emerge as potential solution.

Hamid et al. [8] reduce the features to achieve the same or better result after ranking and reduce the processing time for classification. Classification to recognize the handwritten which is SVM, KNN and Neural Network. Bayes Net classifier is the best classifier with 100% correct for priority time, speed and relations and for FAR (False Acceptance Rate) of 3.13%.

Berchmans et al. [9] proposed that OCR improves efficiency of character Recognition. The techniques for enhancing the quality of the images are character segmentation, character recognition and digital dictionaries. Substantial researches are being carried out to develop a system with intelligence to recognize the natural handwriting with minimal errors in almost all the scripts around the world. More sophisticated OCRs are available for Japanese, Chinese Arabic and Roman primer.

Desai et al. [10] proposed that To Recognize Handwritten Numeral Optical Character using Neural Network. This work approximately 82% of success rate for Gujarati handwritten digit identification.

Prasad et al. [11] proposed that An Adaptive Neuro Fuzzy Classifier (ANFC) for recognition of isolated handwritten characters of Gujarati based on. Fuzzy classification is the task of partitioning a feature space into fuzzy classes. Ensure first that ANFC can be designed to handle large vocabulary problems however recognition rates are not that promising. From the experiment results, it is summarized that the adaptively adjusted classifier performs well on character recognition problem. Drawback of ANFC is eliminated to some extent by ANFC-FH.

Mehta et al. [12] Good recognition rate for Indian scripts, but this rate is for an individual character. Character with different modifiers and specially connected or joint characters are quite complex to identify.

4. CONCLUSIONS

Development of a good OCR system helps people in avoiding the manual entry of relevant documents when entering them into electronic databases. Many researchers have achieved

good recognition rate for Indian scripts, but this rate is for an individual characters. Character with different modifiers and specially connected or joint characters are quite complex to identify. No such OCR system commercially available for Gujarati script that makes digitized printed or and written text searchable with 100% accuracy. It is due to the freestyle writings of the individual and varies from person to person. Due to stumble writing or laziness, it is needless to say that the same writer's specimens on a particular character may differ at the various instances. Moreover, handwriting of different individuals varies because it influenced by many factors such as received education for writing, the quality of paper, the printing materials used, and other factors like stress, motivation and even the purpose of the handwriting. For the last two decades, researchers have been working hard in the field of Gujarati character recognition. Despite everything, it remains an open problem in transforming the document into its digitized form even though sophisticated devices scanners, PDAs are available.

5. REFERENCES

[1] K. S. Siddharth, M. Jangid, R. Dhir, and R. Rani, "Handwritten Gurmukhi Character Recognition Using Statistical and Background Directional Distribution Features," vol. 3, no. 6, pp. 2332–2345, 2011.

[2] M. V. Beigi, "Handwritten Character Recognition Using BP NN , LAMSTAR NN and SVM," 2015.

[3] S. G. Trivedi and A. Nandurbarkar, "Offline Handwritten Character Recognition for Gujarati Language," no. May, pp. 136–139, 2017.

[4] S. J. Macwan, "Classification of Offline Gujarati Handwritten Characters," pp. 1535–1541, 2015.

[5] G. Mehul, P. Ankita, D. Namrata, G. Rahul, and S. Sheth, "Text-Based Image Segmentation Methodology," Procedia Technol., vol. 14, pp. 465–472, 2014.

[6] J. R. Prasad, "Image Normalization and Preprocessing for Gujarati Character Recognition," vol. 3, no. 5, pp. 334–339, 2014.

[7] A. Sharma, P. Thakkar, D. M. Adhyaru, and T. H. Zaveri, "Features Fusion based Approach for Handwritten Gujarati Character Recognition," vol. 5, 2016.

[8] N. Nur, B. Amir, and A. Hamid, "Handwritten Recognition Using SVM , KNN and Neural Network dependent and independent features of text in the process of."

[9] D. Berchmans, "Optical Character Recognition?: An Overview and an Insight," pp. 1361–1365, 2014.

[10] A. A. Desai, "Gujarati handwritten numeral optical character reorganization through neural network," Pattern Recognit., vol. 43, no. 7, pp. 2582–2589, 2010.

[11] J. R. Prasad and U. V. Kulkarni, "Gujrati Character Recognition Using Adaptive Neuro Fuzzy Classifier," 2014 Int. Conf. Electron. Syst. Signal Process. Comput. Technol., pp. 402–407, 2014.

[12] N. Mehta, "A Review of Handwritten Character Recognition," vol. 165, no. 4, pp. 37–40, 2017.

Table I: A Survey on Character Recognition

Publication/Year	Title	Overview	Positive Aspects	Disadvantage
IEEE/2017	Offline Handwritten Character Recognition for Gujarati Language[3]	Systems classified according to data acquisition techniques(Online & Offline) and Systems classified according to the text type(Printed & Handwritten).	Recognizing Gujarati Characters from Scanned Image.	Data samples of Handwritten Gujarati Characters from different writers on plain white papers.
IEEE/2015	Classification of Offline Gujarati Handwritten Characters[4]	Different algorithms from different domains have been considered for comparative analysis like Transform Domain (DWT, DCT and DFT), from Spatial Domain; Geometric Method (Gradient feature), Structural method (Freeman chain code) and Statistical method (Zernike Moments).	Recognizing Gujarati Characters from Scanned Handwritten numeric character Image.	Less storage space needed.
ELSEVIER/2014	Text based Image Segmentation Methodology[5]	First, the need for segmentation is justified in the context of text based information retrieval. Then, the various factors affecting the segmentation process are discussed. Followed by the levels of text segmentation are explored. Finally, the available techniques with their superiorities and weaknesses are reviewed, along with directions for quick referral	To segment a text based image	Only for the printed text document.

		are suggested.		
IEEE/2014	Image Normalization and Preprocessing for Gujarati Character Recognition[6]	Moment based image normalization applied in the pre-prepossessing stage of a character recognition task.	A moment based normalization process for character images for the purpose of enhancing the performance of character recognize for isolated characters of Gujarati.	only for the handwritten text document.
NU/2016	Features Fusion based Approach for Handwritten Gujarati Character Recognition[7]	Recognition of isolated Gujarati handwritten characters is proposed using three different kinds of features and their fusion. Chain code based, zone based and projection profiles based features are utilized as individual features.	To Extract character using Feature Fusion	Increase in number of classes can make the problem difficult and fusion of some more significant features may emerge as potential solution.
IEEE/2016	Handwritten Recognition Using SVM, KNN and Neural Network[8]	Classification to recognize the handwritten which is SVM, KNN and Neural Network.	Reduce the features to achieve the same or better result after ranking and reduce the processing time for classification.	Only for handwritten number
IEEE/2014	Optical Character Recognition: An Overview and an Insight[9]	OCR improves efficiency of character recognition	The techniques for enhancing the quality of the image, character segmentation, character recognition and digital dictionaries.	More sophisticated OCRs are available for Japanese, Chinese Arabic and Roman primer
ELSEVIER/2014	Gujarati handwritten numeral optical character reorganization through neural network[10]	To Recognize Handwritten Numeral Optical Character using Neural Network	To Recognize Handwritten Numeral Optical Character	Only for Gujarati Handwritten Numeral character
IEEE/2014	Gujarati Character Recognition using Adaptive Neuro Fuzzy Classifier[11]	An Adaptive Neuro Fuzzy Classifier (ANFC) for recognition of isolated handwritten characters of Gujarati based on. Compare the performance of ANFC with weighted classifier proposed in by them. Fuzzy classification is the task of partitioning a feature space into fuzzy classes.	Evaluates performance of some efficient classifiers for handwritten characters of Gujarati	Drawback of ANFC is eliminated to some extent by ANFC-FH.
IEEE/2017	A Review of Handwritten Character Recognition[12]	OCR is a machine learning process in which a machine is made to read like human.	Recognizing Gujarati Characters from Scanned Image.	Character with different modifiers and specially connected or joint characters are quite complex to identify