DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4791

ISSN No. 0976-5697

Volume 8, No. 8, September-October 2017

International Journal of Advanced Research in Computer Science

SURVEY REPORT

Available Online at www.ijarcs.info

A TRANSITORY SURVEY OF TOPICAL TRENDS IN INDIC HANDWRITTEN CHARACTERS RECOGNITION

Gautam
Department of Computer Science & Engineering
Pondicherry University
Puducherry, India

K. Vaitheki
Department of Computer Science & Engineering
Pondicherry University
Puducherry, India

Abstract: - Handwritten Recognition, under pattern recognition, is a field having a diverse perspective in the present world scenario. A variety of handwritten text are needed to be recognized given the large quantity of offline and hard copy files need to be converted into a digitized format. An example of offline handwritten character recognition includes documents for office files, bank cheques, important reports and criminal and civil records files. However, for most of the languages that are common, i.e. English, it is particularly easy to convert images into textual data rather than any other scripts, which are complex. Of all the complex scripts, a particular script is known as Indic Scripts, which contains various scripts such a Devanagari, Bangla, Gurumukhi, Dravidian and such other scripts. In this paper, we present a survey of various Indic scripts and its recognition with respect to their corresponding approaches. We make a survey and present the comparative accuracy of several scripts belonging to Indic scripts.

Keywords: Handwritten Character Recognition, Pre-processing, Feature Extraction, Classification, Zone segmentation

I. INTRODUCTION

In the present world, Handwritten Character Recognition [1] becomes a part of one of the important daily applications. Such applications include recognizing a word on a street board, or identifying a text from a piece of paper, or recognizing a text from a camera. All of these approaches are the example of online handwritten character recognition. However, in the offline handwritten character recognition, the scanned or an photographic image of a handwritten text is captured and is embedded into the system which identifies and recognizes the handwritten text, which includes documents such as files, bank cheques and criminal and civil records. In the digital era, where everything is digitized, it is a bizarre need for converting all the old documents, which are present in the records to a digitized format. Now, in an era which digitization is merely not only a push, but also a need, it is high time to move towards a digital era, not only form the scratch but also taking documents from the past and convert into a digitized format.

However, converting a handwritten text to textual format is not an easy task. A handful of algorithms is applied at several stages to convert a handwritten text to a textual format. Whenever a handwritten character has to be converted into textual form, several scripts have to be converted. These scripts are so much independent of each other such that a same algorithm applied on two different scripts gives different results.

Various researches and methodologies have been carried out in the past to determine and improve accuracies of characters and word given in an image. However, the accuracies for the model preferably depends on the type of scripts. Mostly, the work of recognizing an image of a handwritten characters, word, digits, symbols, etc depends upon the type of algorithms used in various phases.

In case of English, Indic and other scripts, recognition of a word, or a character is difficult because the writing style in both the characters is different Moreover, the written text depends upon the type of stroke, orientation of the character, and the writer's way of writing a character. All these factors have an effect on the accuracies, as a factor for a script may not be a factor for the other. All the scripts are independent and a method on two scripts applied may have different accuracies.

Section II discusses various steps involved in recognizing a Handwritten Character. Section IV describes the difference between English and Indic Handwritten scripts. Section IV describes the work done on the recent developments in Indic Handwritten Character Recognition and hence we conclude the survey in Section IV.

II. STEPS IN HANDWRITTEN CHARACTER RECOGNITION

There are different steps involved in the recognition of character, which is mentioned in Fig 1. The characters, in order to be recognized is to be preprocessed well such that the features are extracted in proper. The features hereby extracted from the character are used to train the classifier. From the trained datasets, test images is fed into the classifier to classify the character using the features it takes. The different steps involved in Handwritten Character Recognition is given below:

A. Preprocessing

The first stage, i.e. the preprocessing stage includes some initial processing of an image of a character, before it is given to the classifier. The steps include converting an image of a text into a binary format, if the system is taking binary image as an input. The next step includes segmenting a text first line by line and then word by word. Both line-by-line segmentation and word-by-word segmentation includes

different segmentation algorithms. After segmentation, the words are segmented into characters and therefore, every character is converted into a binary format. The binary image of a character is resized and then correction of slantness, skewness and other such operations are performed such as noise reduction. Finally, the preprocessed image of a character is given to the feature extraction phase.

B. Feature Extraction

Subsequent to the preprocessing phase, features are mined from the individual image of characters, for which different feature extraction techniques exist. For the feature extraction of a character, two types of feature extraction techniques exist- spatial and histogram based features. Spatial features involves extraction of features of a character by taking the shape of a character, i.e. Freeman Chain Code [2]. The other technique i.e. Histogram based features includes construction of a histogram of a character image and taking the histogram based features into the classifier for classification. A wide variety of survey has been done on different feature extraction methods [3].

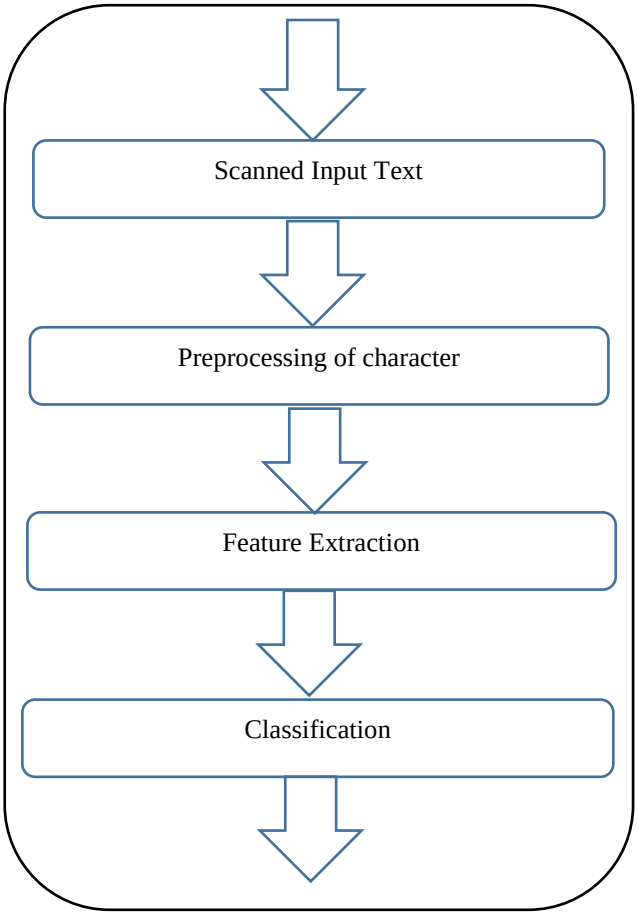

Fig 1. Steps in recognition of a Basic English Handwritten Character

C. Classification

After the feature extraction phase, the extracted features are classified using a classifier. The classifier [4] classifies the input based on the given features and gives an output according to the present training dataset. A classifier needs to be trained with a dataset where a dataset consists of large sample of data whereas to describe the given testing dataset belongs to a particular set of data. The features in the previous phase of feature extraction is done for both training and testing samples. The features of the training samples are extracted and input to the classifier, which constructs a training image. Thereby, for a testing dataset the number of testing samples are tested in the classifier, which tests the testing dataset with a structure formed with the training dataset. There exists a number of classifiers present for the classification of the data of which few are Support Vector Machine, Hidden Markov Model, Neural Networks, K-Nearest Neighbour, etc.

III. INDIC SCRIPT RECOGNITION AND ITS DIFFERNCE WITH INDIC SCRIPT

As far as scripts are concerned, various scripts exists for different languages. A different aspects of research has been carried out for various languages such as Arabic, Chinese, etc. A lot of work has been carried under subsequent research under Arabic and Chinese script recognition. However, in concern with Indic languages such as Devanagari, not a lot of research has been carried out in the years, probable due to the high usage of English as a medium of reading. However, it is important to mark up such aspects of a language as minor sections of the society still rely on it.

A difference between English/digit and Indic say Devanagari handwritten script involves the usage of modifiers. In the absence of modifiers, English/digit or any other such scripts are easy to recognize, as these scripts does not involve use of any modifiers, but the same does not exists for Devanagari characters as it involves the use of modifiers associated with the existing character, which makes a combination of a character and a vowel. A proper difference between the English character and a Devanagari character is given in Fig 2.

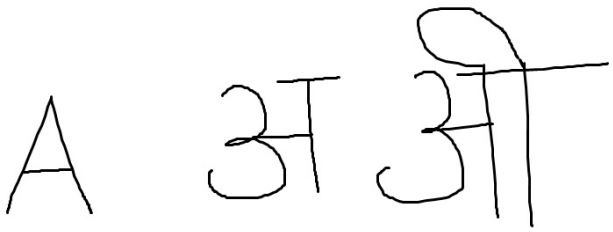

Fig 2. (a) An English character (b) A Devanagari Character with and without Modifier

Fig 2. States the use of a modifier, where modifier usage is not needed in the case of English handwritten character whereas for a Devanagari character, we have a character exists with the modifiers. As we can note that the simple character (a) given in English, with vowels can be modified as per the subsequent characters. But in the case of Devanagari (b), we need to use the modifiers in the given

description of the character to modify a character and attach it with a vowel.

The type of stroke in an English script and Indic scripts is also different in the case that Indic scripts involves multiple strokes and the strokes varies with the types of writers. Therefore, it is not easy to comprehend these strokes, which therefore uses a complex set of algorithms for the recognition of an Indic script.

IV. STUDY OF INDIC SCRIPTS AND ITS RECOGNITION

In the case of Devanagari script recognition, it is important to note that not many research works had been carried out. The number of work on Indic handwritten script recognition is illustrated below and in Table 1.

Ghosh et.al [5] has made a comparative study of three feature extraction techniques for two Indic scripts- Devanagari and Bengali using Hidden Markov Model. In the first place, approach utilizes the entire stroke highlight extraction without neighborhood zone division, while the other two methodologies consider the division of the word into essential strokes and nearby zone investigation on each stroke. In these two nearby zone shrewd highlights, one takes auxiliary and directional highlights while different takes overwhelming focuses identified from slant points, to get neighborhood highlights. From the result, dominant point local feature extraction method works best for the dataset and provide better accuracies for both Bangla and Devanagari scripts with an accuracy of 90.23% and 93.82% for both Bengali and Devanagari scripts respectively.

Roy et.al [6] has testified a novel approach for offline Bangla handwritten word recognition by Hidden Markov Model (HMM). The image of the word is segmented into 3 zones, upper, middle and lower, respectively. Features are extracted from all the three zones separately using Local Histogram Gradient (LGH) [7] dividing the image into 4x4 cells and hence is used for classification. The upper and the lower zones are classified using SVM and the middle zone is classified using HMM. Performance evaluation is performed on the dataset of 10.120 Bangla handwritten words which shows an accuracy of 67.12%.

Pagare et.al [8] proposed an acknowledgment display for digitizing manually written Devanagari characters and an auto affiliated acknowledgment system for Devanagari characters and numerals proposed in the present work by utilizing classifiers. To solve recognition problem a dynamic model based on Hopfield neural network deployed. The average accuracy for this model is 92.91%.

Roy. et.al [9] projected a segmentation-centered approach for Devanagari handwritten word recognition. The work divides the word into three zones: upper zone, middle zone, lower zone. For Matra detection, the work proposes a water reservoir concept. The work illustrates the use of Hidden Markov Model [10] for middle zone recognition and Support Vector Machine [11] for Upper and Middle Zone recognition. The entire work includes the reconciliation of water repository perception [12] for improved zone division in a word picture, a proficient PHOG [13] highlights created to enhance the execution of HMM based center zone recognition [11]. The proposed structure has been summed up and tried for Bangla and Devanagari contents acknowledgment. However according to Devanagari content, the exactness for the acknowledgment is 84.24% and 94.51% precision with top 1 and best 5 decisions.

Pal et.al [14] proposed an algorithm to recognize words that are oriented or that are curved in shape for Devanagari and Bangla words. If there should arise an occurrence of Bangla and Devanagari contents, characters more often than not frame word touch and hole districts which is taken care by the foundation data of such word is utilized. Here, convex hull and water repository standard have been connected. The characters are segmented from the documents using the background information of the word. Henceforth, individual characters are recognized using rotation invariant features from the foreground part of the characters. The touching part is recognized utilizing water supply standard. Contingent upon composing mode and the repository base-locale of the touching part, an arrangement of applicant envelope focuses is then chosen from the form purposes of the segment. In view of the competitor focuses, the touching part is at long last fragmented into singular characters. Consequences of 99.18% (98.86%) precision when tried on 7515 (7874) Devanagari (Bangla) characters.

Table 1. Present Methods and its accuracy

Sl.no	Proposed Method	Accuracy
1.	Comparison of Zone-Features for Online Bengali and Devanagari Word Recognition using HMM [5]	90.23%(Bangla) 93.82 %(Devanagari)
2.	A Novel Approach of Bangla Handwritten Text Recognition using HMM [6]	67.12%(Bangla)
3.	Associative Memory Model for Distorted On-line Devanagari Character Recognition [8]	92.91%(Devanagari)
4.	HMM-based Indic handwritten word recognition using Zone segmentation [9]	84.24% and 94.51% (for Bangla scripts)
5.	Multi-oriented Bangla and Devanagari text recognition [14]	98.86%(Bangla)
6.	Gujarati handwritten numeral optical character reorganization through neural network [15]	82%(Gujarati)
7.	HMM-Based Lexicon-Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts [19]	87.13%(Devanagari) 91.8%(Tamil)

Desai [15] proposed a work involving recognition of Gujarati digits. The method includes the use of neural networks [16] [17] [18] for the recognition of a Gujarati digit, which is performed using multi layered feed forward

neural network. Preprocessing involves thinning, contrast adjustment using Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm and skewness correction with feature extraction, which includes the use of four different profiles, horizontal, vertical, and two diagonals, i.e. four different profiles of digits(X-profile, Y-profile, diagonal 1 profile and diagonal 2 profile). The accuracy rate for this work has reached 82%.

Bharath et.al [19] proposed a solution for two major Indic scripts-Tamil and Devanagari, as largely script independent and data driven. Two major techniques are proposed: lexicon free and lexicon driven based on Hidden Markov Model (HMM). Lexicon driven technique maps each word in lexicon as a classification of HMM representation symbols. The lexicon free technique uses a handwritten word with a Bag of symbols representation independent of nonstandard symbol writing orders, which allows lexicon pruning. The pre-processing of script includes size normalisation and resampling, added to Devanagari where to detect the stroke, i.e. shirorekha detection(The horizontal stroke of the character or a word) which is done using Shirorekha Detection Algorithm as described by Joshi et.al[20].After pre-processing, a set of 9 features are extracted from NPen++ recognition algorithm[21]. Hereafter, the symbols are modelled using Hidden Markov Model and lexicon matching is performed using lexicon driven matching and lexicon free matching. 20,000 word lexicons were used where Devanagari uses the combination of lexicon free and lexicon independent technique giving an accuracy of 87.13% whereas Tamil using only lexicon driven technique providing an accuracy of 91.8%.

V. CONCLUSION

This paper investigates the recent trends and research in Indic script recognition. It is very essential to have a better accuracy rate such that it can be applied for the real world applications. In the case of accuracy, English scripts have a lot of research work undergone by various researchers, providing a higher accuracy rate. In addition, it is applied in real world applications where the use of English scripts is most dominant. However, for other scripts, which are not prominent in use, such as Indic scripts, not many research works have been carried out. The fact that English scripts differ from Indic scripts which includes various other scripts such that it involves the use of modifiers, which makes it difficult to recognize. Also, the fact that Indic scripts have a diverse range of scripts, Devanagari, Bangla, Gurumukhi, Dravidian to name a few are different from each other, whatsoever few similarities may still exist. The rate of accuracy until now for all the scripts are well and acceptable but still various problems still exist in the present methods whatsoever the accuracy rate may be. The recognition of modifiers along with the middle zone of a character, in few of the cases is one of the problem to name a few in the case of Devanagari and Bangla scripts recognition. Unlike, in case of English scripts where the segmentation is an easier method without involving the use of modifiers, Indic scripts involves the use of modifiers, which makes the recognition task very tedious. Such cases includes the recognizing modifiers and middle zone separately and then classifying it separately.

Throughout the survey, it is mentioned that not many major research works have been done on Indic Handwritten script recognition. Hence, more number of Indian scripts can be tried as an input for the existing methods along with improving accuracy of the existing methods. Works including scripts with numbers can also be carried out as a scope of work. Thus, a need of digitizing must be carried over by including new works specifically on the Indian scripts by introducing a large number of major research works.

VI. REFERENCES

[1] Lphabets, a. (2016). A survey on handwritten character recognition (hcr) techniques for english, *3*(1).

[2] Azmi, A. N., & Nasien, D. (2014). Feature vector of binary image using Freeman Chain Code (FCC) representation based on structural classifier. *International Journal of Advances in Soft Computing and Its Applications*, *6*(2), 1–19.

[3] Gunawan, F. E., Hapsari, I. A., Soewito, B., & Candra, S. (2016). A Study of Comparison of Feature Extraction Methods for Handwriting Recognition, 73–78.

[4] Technology, I. (n.d.). Comparative Study of Devanagari Handwritten and printed Character & Numerals Recognition using Nearest-Neighbor Classifiers.

[5] Ghosh, R., & Roy, P. P. (2017). Comparison of zone-features for online Bengali and Devanagari word recognition using HMM. *Proceedings of International Conference on Frontiers in Handwriting Recognition, ICFHR*, 435–440. https://doi.org/10.1109/ICFHR.2016.0087

[6] Roy, P. P., Dey, P., Roy, S., Pal, U., & Kimura, F. (2014). A Novel Approach of Bangla Handwritten Text Recognition Using HMM. *2014 14th International Conference on Frontiers in Handwriting Recognition*, 661–666. https://doi.org/10.1109/ICFHR.2014.116

[7] Rodríguez-Serrano, J. A., & Perronnin, F. (2009). Handwritten word-spotting using hidden Markov models and universal vocabularies. *Pattern Recognition*, *42*(9), 2106–2116. https://doi.org/10.1016/j.patcog.2009.02.005

[8] Pagare, G., & Verma, K. (2016). Associative Memory Model for Distorted On-Line Devanagari Character Recognition. *Proceedings - 2015 5th International Conference on Advances in Computing and Communications, ICACC 2015*, 46–49. https://doi.org/10.1109/ICACC.2015.42

[9] Roy, P. P., Bhunia, A. K., Das, A., Dey, P., & Pal, U. (2016). HMM-based Indic Handwritten Word Recognition using Zone Segmentation Author ' s Accepted Manuscript. *Pattern Recognition*, *60*(May), 1–31. http://doi.org/10.1016/j.patcog.2016.04.012

[10] Procter, S., Illingworth, J., & Mokhtarian, F. (2000). Cursive handwriting recognition using hidden Markov models and a lexicon-driven level building algorithm. *IEE Proceedings - Vision, Image, and Signal Processing*, *147*(4), 332. https://doi.org/10.1049/ip-vis:20000476

[11] Gruber, C., Gruber, T., Krinninger, S., & Sick, B. (2010). Machines Based on LCSS Kernel Functions, *40*(4), 1088–1100.

[12] Roy, R. K. (2012). Multi-lingual City Name Recognition for Indian Postal Automation, (1). https://doi.org/10.1109/ICFHR.2012.238

[13] Bai, Y., Guo, L., Jin, L., & Huang, Q. (2009). A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition. *Proceedings - International Conference on Image Processing, ICIP*, (7118074), 3305–3308. https://doi.org/10.1109/ICIP.2009.5413938

[14] Pal, U., Pratim Roy, P., Tripathy, N., & Llads, J. (2010). Multi-oriented Bangla and Devnagari text recognition. *Pattern Recognition*, *43*(12), 4124–4136. https://doi.org/10.1016/j.patcog.2010.06.017

[15] Ã, A. A. D. (2010). Gujarati handwritten numeral optical character reorganization through neural network, *43*, 2582–2589. https://doi.org/10.1016/j.patcog.2010.01.008

[16] C. Luh Tan, A. Juntan, Digit recognition using neural networks, Malaysian Journal of Computer Science 17 (2) (2004) 40–54.

[17] M.B. Sukhswami, P. Seetharamulu, A. Pujari, Recognition of Telugu characters using neural networks, International Journal of Neural Systems 6 (3) (1995) 317–357

[18] M. Wellner, J. Luan, C. Sylvestor, Recognition of Handwritten Digits using Neural Network, /http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.136.9800S.

[19] Bharath, A., Madhvanath, S., & Member, S. (2012). HMM-Based Lexicon-Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts, *34*(4), 670–682.

[20] N. Joshi, G. Sita, A.G. Ramakrishnan, V. Deepu, and S. Madhvanath, "Machine Recognition of Online Handwritten Devanagari Characters," Proc. Eighth Int'l Conf. Document Analysis and Recognition, pp. 1156-1160, Aug.-Sept. 2005.

[21] S. Jaeger, S. Manke, J. Reichert, and A. Waibel, "Online Handwriting Recognition: The NPen++ Recognizer," Int'l J. Document Analysis and Recognition, vol. 3, no. 3, pp. 169-180, Mar. 2001.