HWR for Indian Languages: A Comprehensive Survey

Jino P J
Department of Computer Applications
CUSAT
Cochin,India
jinopallickal@gmail.com

Kannan Balakrishnan
Department of Computer Applications
CUSAT
Cochin,India
bkannan@cusat.ac.in

***Abstract*— Few major Research works are going in the field of Handwriting Word Recognition (HWR) of Indian languages. This paper surveys the major works of offline/online handwritten word recognition. Techniques involved in word recognition are also discussed. Major works carried out in Bangla, Urdu, Tamil and Hindi are mentioned in this paper. Advancement towards HWR in other Indian languages are also discussed. Application of offline HWR is also discussed.**

***Index Terms*—HWR, online, offline, applications**

I. Introduction

Handwriting Word Recognition (HWR) involves the conversion of handwritten text on an image into a compute readable format. In the case of online writing which pertains to the availability of trajectory data during writing. There are 10 major scripts in India for the documentation of its official languages. They are Devanagari, Bangla, Gurumukhi, Guajarati, Oriya, Kannada, Telugu, Tamil, Malayalam and Urdu (Nastaliq)[1]. The research on HWR aims at the development of software products capable of processing the images of the paper documents with different scripts and writing styles, and also interpreting the text written by the user. The organization of the survey is as follows. Section 2 of the paper gives about a brief overview about the features of Indian Languages. Section 3 of the paper covers the techniques used for HWR in Indian regional scripts. Major HWR works are done in the four scripts are Hindi, Bangla, Urdu and Tamil. Advancements towards the development of HWR in other Indian languages are also discussed in this section. Fourth section discusses the application of handwritten regional scripts. Last section concludes the survey.

II. Features of Indian Language

Indian Scripts are rich in patterns while the combinations of such patterns makes the problem even more complex. Some of the features of Indian languages and the scripts used to express them are [2] :

A. *Phoneme Set*

Indian languages have a more sophisticated notion of a character unit or akshara that forms the fundamental linguistic unit. An akshara consists of 0, 1, 2, or 3 consonants and a vowel. Words are made up of one or more aksharas. Each akshara can be pronounced independently as the languages are completely phonetic. Aksharas with more than one consonant are called samyuktaksharas or combo-characters. The last of the consonants is the main one in a samyuktakshara.

All Indian languages have essentially the same alphabet derived from the Sanskrit alphabet. This common alphabet contains 33 consonants and 15 vowels in common practice. Additional 3- 4 consonants and 2-3 vowels are used in specific languages or in the classical forms of others. This difference is not very significant in practice. Individual consonants and vowels form the basic letters of the alphabet.

B. *Different Grapheme's*

The commonality in the alphabet does not extend the graphic forms used to express them in print. Each language uses different scripts consisting of dissimilar grapheme's for printing. Thus, printed matter in other scripts are inaccessible to readers of one script. There are 10 major scripts in India. The Devanagari script is the widest used one, being used to write Hindi (the most spoken language), Marathi, Konkani, and Nepali, the language of the neighboring Nepal. Different scripts use different philosophies for the individual grapheme's and their combinations. Some have a head-line or shirorekha that persists for a whole word. Others have non-touching grapheme's. The grapheme of one of the consonants is usually at the heart of the printed akshara. The vowel appears as a matra or vowel modifier. These can appear to the left,right, above or below it or in combinations. The supporting consonants of a samyuktashara also appear as modifier grapheme's to the left, right, above, or below of the main one. These modifiers could be truncated or scaled down forms of the basic consonant, but could also be completely different. They may touch each other or the main consonant in some cases or may be separate. These rules are not consistent even within a script and certainly not across scripts.

C. *Reduplication*

All languages employ reduplicated form in varying degrees and for different functions, extensive use of reduplication is a particular characteristic of Indian Languages. different kinds of re duplicative expressions found in the Indian Laguages are Onomatoeic, Expressive, Paired Words, Echo.

III. Techniques for HWR In Indian Regional Scripts

A word recognition algorithm attempts to associate the word image to choices in a lexicon [3]. Ranking is produced using this algorithm. This is done either by the analytic approach of recognizing the individual characters or by the holistic approach of dealing with the entire word image. The latter approach is useful in the case of touching printed characters and handwriting. A higher level of performance is observed by combining the results of both approaches [4]. There exist several different approaches to word recognition using a limited vocabulary [5].

A. *Advancements in HWR of Bangla Script*

A.1 Recognition of Handwritten Words Using Neural Network Based Techniques.

Basu et al. [2009] [6]presented a hierarchical approach for the recognition of handwritten Bangla words. The approach segmented a word image on headerline hierarchy, then recognizes the individual word segments and then identified the constituent characters of the word through intelligent combination of recognition decisions of the associated word segments. MLP-based pattern classifiers are used in the work for most of the classification tasks. The three types topological features considered here are longest run features, modified shadow features, and octant-centroid features.

Table 1: Major Works in Bangla Script

Methodology	Features	Classifier	Data set ((Size)	Accuracy (%)
Pal et al. [2009][7]	Directional	DP, MQDF	8,625	94.08
Vajda and Belaid [2005][8], Vajda et al. [2009]	Low & High level	NSHP-HMM	7,500	86.8
Basu et al. [2009][6]	Topological	MLP	127	80.58
Bhowmik et al. [2008][9] [2012](Lexicon Reduction Technique)[10]	Shape based directional	HMM & GA	35,700	

A.2 MQDF Based Techniques for Recognizing Handwritten Words.

Pal et al. [2009] proposed a lexicon driven segmentation based recognition scheme for Bangla handwritten city name recognition[7]. A water reservoir concept was applied to segment the words into possible primitive Components. These components were then merged into possible characters to get the best match using the lexicon information. To merge these primitive components into characters, dynamic programming (DP) was applied using total likelihood of the characters of a city-name as the objective function. To compute the likelihood of a character, Modified Quadratic Discriminant Function (MQDF) was used. The features used in the MQDF were the directional features of the contour points of the components.

A.3 HMM Based Techniques for Recognizing Handwritten Words.

Proper segmentation of characters from handwritten words is a difficult task because of various writing styles. Because of these segmentation problems, Vajda and Belaid [2005][8] and Vajda et al. [2009] proposed a context based, segmentation-free Hidden Markov Model (HMM) recognition system for handwritten Bangla words. The approach combined a Markov Random Field (MRF) and a Non-Symmetric Half Plane Hidden Markov Model (NSHPHMM). Low-level pixel information and high level structural features were combined in the framework of NSHP-HMM. Here information coming from the structural nature of the pixels allowed researchers to precisely measure the quantity and quality of the information perceived by the HMM and that helped to improve the results. Bhowmik et al. [2008] proposed a recognition system for isolated handwritten Bangla city names (fixed lexicon) using a left-right Hidden Markov Model (HMM)[9]. A genetic algorithm (GA) was used to train the HMM with shape-based direction encoding features.

B. *Advancements in Urdu HWR*

Urdu is written from right to left and is an extension of the Persian alphabet (Farsi). Urdu was mainly developed in the Uttar Pradesh state of the Indian subcontinent, but began taking shape during the Delhi Sultanate as well as the Mughal Empire (1526–1858 AD) in South Asia. Standard Urdu is conventionally written in Nastaliq calligraphy style having 38 characters. Vowels in Urdu are represented by letters, which are also considered as consonants. In India, Urdu is also an official language for documentation.

For holistic recognition of handwritten Urdu words, Sagheer et al. [2010] used structural and gradient (SG) features along with a support vector machine (SVM) for classification[11]. Previously, Sagheer et al. [2009] only used gradient features for handwritten Urdu numeral recognition using SVM[12]. In the scheme proposed by Mukhtar et al. [2010] for the classification and recognition of handwritten Urdu words, gradient, structural, and cavity (GSC) features were used along with a support vector machine (SVM).

Table 2: Major works in Urdu HWR

Methodology	**Features**	**Classifier**	**Data set ((Size)**	**Accuracy (%)**
Sagheer et al. [2010][11]	SG	SVM	19,432	97
Mukhtar et al. [2010][13]	GSC	SVM	1,300	70

C. *Advancement in Tamil HWR*

Tamil is an ancient, classical Dravidian language in existence for over two thousand years. The Tamil script traces its roots to the Brahmin script and continues t to undergo a lot of changes to transgress itself as a portable medium. There are 30 basic shapes (12 vowels and 18 consonan nts) in Tamil. In addition to this there are six Grantha characters and one Ayutham**.** With the combination of these 30 basic shapes and one ayutham letter, in total we get two hundred and forty seven characters that are used in Tamil language.

For holistic recognition of handwritten Tamil words, Thadchanamoorthy Subramaniam et al.[2012] used Gabor features along with other geometric features of the word imag ge are then fed to an SVM classifier for recognition[14].

In another work Thadchanamoorthy et al[2013] developed a city name data base for the postal automation[15].Here the recognition is based upon the lexicons, so proper segmentation is not required. Here the binarized city names are pre segmented to individual character or parts. Then these parts are merged into a city name using dynamic programming.

Bharath et al[2007] proposed a data driven Hidden Markov Model for Online Handwritten Tamil Word Recognition[16].

Table 3: Major works in Tamil HWR

Methodology	**Features**	**Classifier**	**Data set ((Size)**	**Accuracy (%)**
Thadchanamoorthy et al[2012] [14]	Gabour	SVM	4270	86.36
Thadchanamoorthy et al[2013] [15]	Direction		265	96.89
Bharath et al[2007][16]		HMM	1000 20000	98 92.2

D. *Advancement in Hindi HWR*

Words in Indic languages are composed of a number of aksharas or syllables, which in turn are formed by groups of consonants and vowel modifiers. Segmentation of aksharas is critical to accurate recognition of both recognition primitives as well as the complete word. Also, recognition in itself is a complex job.

Holistic Recognition of online handwritten isolated Hindi words Belhe et al[2013] used a combination of HMMs trained on Devanagari symbols and a tree formed by the multiple, possible sequences of recognized symbols[17].

For Offline Handwritten Word Recognition in Hindi Sitaram Ramachandrula et al[2012] used two-pass Dynamic Programming algorithm to match the test word against each word in the lexicon by initially segmenting the test word image into probable characters[18]. In this work they extract directional element features (DEF) on each character image segment and statistically model them. Also they created a Hindi handwritten word and character database from 100 writers for the purpose of training and testing the offline HWR.

Table 4: Major Works in Hindi HWR

Methodology	**Features**	**Classifier**	**Data set ((Size)**	**Accuracy (%)**
Swapni Belhe et al[2012][17]	HOG	HMM	10000	89
Sitaram Ramachandrula et al[2012][18]	DEF	Dynamic Programming Algorithm	10 to 30	91.23 - 79.94

E. *Advancement in Other Indian Languages towards HWR*

In some Indian languages word recognition is in infant Stage. Major works doing in this area that can extend to HWR are listed on Table 5.

Table 5 : Advancement towards HWR in other Indian Languages

Metho-dology	**Language/ Model**	**Features**	**Classifier**	**Data set ((Size)**	**Acc-uracy (%)**
Gupta et al[2013] [19] Agarwalet al[2012] [20]	Gurumukhi /online Stroke level	-	SVM		98.24
Munish Kumar et al[2014] [21]	Gurumukhi /offline Segmentation	-	-		93.51
Patel et al[2013] [22]	Gujarathi/ Segementa-tion	-	-	-	-
M Vishwas et al[2012] [23]	Kannada/ Character Recogntion	The features that are used to form a signature	Combined Direction based Stroke Density principle(	49 Characte-rs & 10 numerals	94.4

		are direction of the stroke, density of the stroke and number of clicks for the character.	DSD) with Kohonen Neural Network (KNN).		
Soman et al[2013] [24]	Telugu Character Recognition		Convolutional neural networks (CNN), Principal Component Analysis (PCA), Support vector machines and Multiclassifier systems.	Numeric data Consonanats Vowels	98.5 92.26 92
Vidya et al[2013] [25]	Malayalam Character Recognition	-	Probabilistic Simplified Fuzzy ARTMAP	-	-
Jomy John et al[2013] [26]	Malayalam Character Recognition	Gradient, curvature calculation and dimensionality reduction	SVM	-	97.96

IV. APPLICATIONS OF OFF-LINE HWR

There has been significant growth in the application of off-line handwriting recognition during the past decade. The most important of these has been in reading postal addresses, bank check amounts, forms and medical treatments.

A. Handwritten Address Interpretation

The task of interpreting handwritten addresses is one of assigning a mail piece image to a delivery address. An address for the purpose of physical mail delivery involves determining the country, state, city, post office, street,primary number (which could be a street number or a post office box), secondary number (such as an apartment or suite number), and finally, the firm name or personal name [27].

A Handwritten Address Interpretation (HWAI) system uses knowledge of the postal domain in the recognition of handwritten addresses. The task is considered to be one of interpretation rather than recognition since the goal is to assign the address to its correct destination irrespective of incomplete or contradictory information present in the writing or due to climate conditions like heavy rains etc..

B. Bank Check Recognition

Bank check recognition presents several research challenges in the area of document analysis and recognition. The backgrounds are often colored and have complex patterns. The type and position of preprinted information fields as well as the guides that prompt patron information vary widely [28]. The handwritten components that are provided by the patron are: 1) legal (worded) amount, 2) courtesy (numeric) amount, 3) date, and 4) the signature [29].

Field layout analysis involves image filtering and binarization, segmentation of text blocks, and removal of guide lines and noise. A complete bank check recognition system, including the layout analysis and recognition components, that are engineered for industrial applications is described in[19] . Hidden Markov Models are used for the recognition of both the legal and courtesy amounts in [30][31].

C. Signature Verification

In a typical off-line signature verification system, a signature image, as scanned and extracted from a bill, a check or any official document, is compared with a few signature references provided, for example, by a user at the opening of his account. Opposite to on-line systems, there is no time information directly available and the verification process relies on the features that can be extracted from the luminance of the trace only. Although the extraction of a signature from a document background is already a very difficult problem in itself, particularly for checks (see, for example, [32]), most of the studies published to date assume that an almost perfect extraction has been done. In other words, the signature specimens used in these studies are generally written on a white sheet of paper.

D. Writer Identification

Handwriting identification deals with comparing questioned writing with known writing exemplars and determining whether the questioned documents and exemplars were written by the same or different authors.

Two issues of concern in this procedure are the variability of handwriting within individuals, which are individual characteristics and between individuals, which are class characteristics. The extraction of distinctive individual traits is what is relied on to determine the author of the questioned document. Information about these two classes of variability is gathered based on the features for characterizing handwriting [33]. Some of the elements of comparison are: alignment (reference lines), angles, arrangement (margins, spacing), connecting strokes (ligatures and hiatuses),curves, form (round, angular or eyed), line quality (smooth, jerky), movement, pen lifts, pick-up strokes (leading ligatures), proportion, retrace, skill, slant, spacing, spelling, straight lines, and terminal strokes.

F. *Medical Applications*

HWR can be used to identify the progress of paralysed patients after the treatment. It can be used to understand the stability of the human being.

V. CONCLUSIONS

Handwriting Word Recogntion is a challenging task and it requires a great level of accuracy. Most of the works done in this area achieved more than 80% of accuracy. Techniques used for HWR is script dependent. Application of HWR is enormous. In the South Dravidian Language like Malayalam HWR is in the initial phase.

REFERENCES

[1] Pal, Umapada, Ramachandran Jayadevan, and Nabin Sharma. "Handwriting Recognition in Indian Regional Scripts: A Survey of Offline Techniques." *ACM Transactions on Asian Language Information Processing (TALIP)* 11.1 (2012): 1.

[2] Madhavi Varalwar, Nixon Patel. " Characteristics of Indian Languages" available at "http:// http://www.w3.org/2006/10/SSML/papers/CHARACTERISTICS_OF_INDIAN_LANGUAGES.pdf" on 30/12/2013

[3] Simon, J-C. "Off-line cursive word recognition." *Proceedings of the IEEE* 80.7 (1992): 1150-1161.

[4] Huang, Y. S., and C. Y. Suen. "The behavior-knowledge space method for combination of multiple classifiers." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. INSTITUTE OF ELECTRICAL ENGINEERS INC (IEEE), 1993.

[5] Gilloux, Michel, Jean-Michel Bertille, and Manuel Leroux. "Recognition of handwritten words in a limited dynamic vocabulary." *Proceedings Int. Workshop on Frontiers in Handwriting Recognition*. 1993.

[6] Basu, Subhadip, et al. "A hierarchical approach to recognition of handwritten Bangla characters." *Pattern Recognition* 42.7 (2009): 1467-1484.

[7] Umapada, P. A. L., R. O. Y. Kaushik, and Fumitaka Kimura. "A lexicon-driven handwritten city-name recognition scheme for Indian postal automation." *IEICE transactions on information and systems* 92.5 (2009): 1146-1158.

[8] Vajda, Szilárd, and Abdel Belaïd. "Structural information implant in a context based segmentation-free HMM handwritten word recognition system for latin and Bangla script." *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. IEEE, 2005.

[9] Bhowmik, Tapan Kumar, Swapan K. Parui, and Utpal Roy. "Discriminative HMM training with GA for handwritten word recognition." *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008.

[10] Bhowmik, Tapan Kumar, Utpal Roy, and Swapan K. Parui. "Lexicon Reduction Technique for Bangla Handwritten Word Recognition." *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on*. IEEE, 2012.

[11] Sagheer, Malik Waqas, et al. "Holistic Urdu handwritten word recognition using support vector machine." *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010.

[12] Sagheer, Malik Waqas, et al. "A New Large Urdu Database for Off-Line Handwriting Recognition." *Image Analysis and Processing–ICIAP 2009*. Springer Berlin Heidelberg, 2009. 538-546.

[13] Mukhtar, Omar, Srirangaraj Setlur, and Venu Govindaraju. "Experiments on urdu text recognition." *Guide to OCR for Indic Scripts*. Springer London, 2010. 163-171.

[14] Subramaniam, Thadchanamoorthy, et al. "Holistic recognition of handwritten Tamil words." *Emerging Applications of Information Technology (EAIT), 2012 Third International Conference on*. IEEE, 2012.

[15] Thadchanamoorthy, S., et al. "Tamil Handwritten City Name Database Development and Recognition for Postal Automation." *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013.

[16] Bharath, A. "Hidden Markov Models for online handwritten Tamil word recognition." *Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on*. Vol. 1. IEEE, 2007.

[17] Belhe, Swapnil, et al. "Hindi handwritten word recognition using HMM and symbol tree." *Proceeding of the workshop on Document Analysis and Recognition*. ACM, 2012.

[18] Ramachandrula, Sitaram, Shrang Jain, and Hariharan Ravishankar. "Offline handwritten word recognition in Hindi." *Proceeding of the workshop on Document Analysis and Recognition*. ACM, 2012.

[19] Gupta, Mayank, Nainsi Gupta, and Rahul Agrawal. "Recognition of Online Handwritten Gurmukhi Strokes Using Support Vector Machine." *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*. Springer India, 2013.

[20] Agrawal, Rahul, and R. K. Sharma. "Recognition of online handwritten Gurmukhi strokes using support vector machine." (2012).

[21] Kumar, Munish, M. K. Jindal, and R. K. Sharma. "Segmentation of Isolated and Touching Characters in Offline Handwritten Gurmukhi Script Recognition." (2014).

[22] Patel, Chhaya, and Apurva Desai. "Extraction of Characters and Modifiers from Handwritten Gujarati Words." *International Journal of Computer Applications* 73 (2013).

[23] Vishwaas, M., M. M. Arjun, and R. Dinesh. "Handwritten Kannada character recognition based on Kohonen Neural Network." *Recent Advances in Computing and Software Systems (RACSS), 2012 International Conference on*. IEEE, 2012.

[24] Soman, Soumya T., Ashakranthi Nandigam, and V. Srinivasa Chakravarthy. "An efficient multiclassifier system based on convolutional neural network for offline handwritten Telugu character recognition." *Communications (NCC), 2013 National Conference on*. IEEE, 2013.

[25] Vidya, V., et al. "Malayalam Offline Handwritten Recognition Using Probabilistic Simplified Fuzzy ARTMAP." *Intelligent Informatics*. Springer Berlin Heidelberg, 2013. 273-283.

[26] John, Jomy, and Kannan Balakrishnan. "A System for Offline Recognition of Handwritten Characters in MalayalamScript." *International Journal of Image, Graphics and Signal Processing (IJIGSP)* 5.4 (2013): 53.

[27] Plamondon, Réjean, and Sargur N. Srihari. "Online and off-line handwriting recognition: a comprehensive survey." *Pattern*

Analysis and Machine Intelligence, IEEE Transactions on 22.1 (2000): 63-84.

[28] Cohen, Edward, Jonathan J. Hull, and Sargur N. SriHari. "Understanding handwritten text in a structured environment: determining zip codes from addresses." *International journal of pattern recognition and artificial intelligence* 5.01n02 (1991): 221-264.

[29] Downton, A. C., R. W. S. Tregidgo, and C. G. Leedham. "Hendrawan, Recognition of handwritten British postal addresses." *From Pixels to Features III: Frontiers in Handwriting Recognition (J. Simon and S. Impedovo, eds.)* (1992): 129-144.

[30] Franke, Katrin, and Mario Köppen. "Towards an universal approach to background removal in images of bankchecks." *in Proceedings 6th International Workshop on Frontiers in Handwriting Recognition (IWFHR), Tajon, Korea*. 1998.

[31] Dimauro, Giovanni, et al. "A multi-expert signature verification system for bankcheck processing." *International Journal of Pattern Recognition and Artificial Intelligence* 11.05 (1997): 827-844.

[32] Dimauro, Giovanni, et al. "Automatic bankcheck processing: A new engineered system." *International Journal of Pattern Recognition and Artificial Intelligence* 11.04 (1997): 467-504.

[33] Bradford, Russell R., and Ralph Bradford. *Introduction to Handwriting Examination and Identification*. Nelson-Hall, 1992.