Online Handwriting Recognition for Tamil Script in Recent Past

Sukhdeep Singh*

Assistant Professor in Computer Science, Mata Sahib Kaur Girls College, Talwandi Bhai, Ferozepur, Punjab, India

Abstract – The recent past has observed a rigorous research on optical character recognition and handwriting recognition (HWR) for Chinese-Japanese- Korean, Latin, Arabic and Indic scripts. Handwriting recognition can understand a handwritten text whether it is written on the paper, palm leaves, stone or digital surfaces. In case of offline handwriting recognition, the text is first written on the paper or any other hard surface, then it is scanned to convert it to digital form of text and finally the computer understands the handwritten text. In case of online handwriting recognition mode of HWR, the handwritten characters/words/sentences are understood by the computer while writing on the digital surface of pen tablets, mobiles, computers or handwriting boards. An online handwritten word recognition (OHWR) model has various steps as data collection, preprocessing, segmentation of connected strokes, feature extraction, classification and post processing etc. The present work has been carried out for the Indic script Tamil, which is an abugida script and used Tamils and Tamil speakers in India. The present work presents the recent work done for Tamil handwriting recognition in online handwriting mode. This work has been done for handwriting recognition of Tamil characters and words. The major results represented in this paper have been taken from reputed journals and conferences of pattern recognition, artificial intelligence and handwriting recognition as Pattern Recognition, Pattern Analysis and Machine Intelligence, Pattern Recognition Letters, International Workshop on Frontiers in Handwriting Recognition, International Journal on Document Analysis and Recognition , International Conference on Document Analysis and Recognition and International Conference on Frontiers in Handwriting Recognition etc.

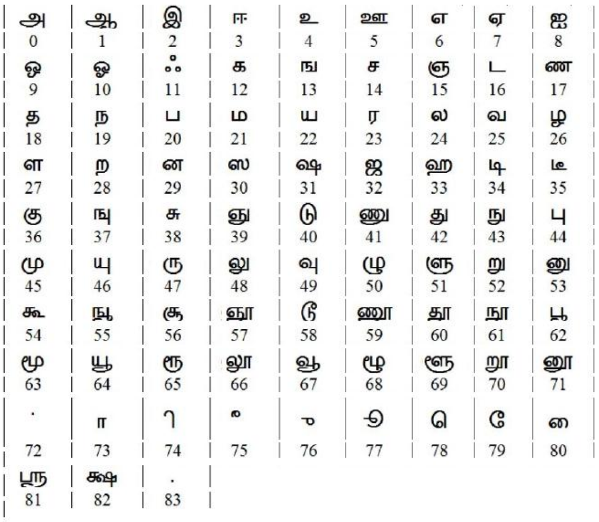
Keywords: Indic, Tamil, Online Handwriting Recognition

- X -

INTRODUCTION

As the time is passing, the technological innovational are also advancing at a rapid rate. It is completely right to say that the today's era is the era of technology. In this modern technological era, the role of machine learning and artificial intelligence cannot be ignored. The modern and advance technology has also played a major role in ways of changing methods to communicate with computers and other computing devices. The handwriting recognition is one the popular ways to communicate with computing devices and also avoids the problems of speech recognition. The handwriting recognition is among the active areas of research for recent decades. Handwriting recognition refers to understanding the handwritten text written on paper or digital surface. When the handwritten text is first written on the paper and then it is scanned and understood by the computer, it is called offline handwriting recognition. But on the other hand, when the handwritten text is understood by the computer while writing on the screen or digital surface of the computer, it is called online handwriting recognition. A lot of work has been done for handwriting recognition in offline and online handwriting mode. A great work has been done for Chinese-Japanese-Korean (CJK), Arabic and Latin scripts. In comparison to these scripts, the available work for offline/online handwriting in Brahmi scripts is less, so it needs the attention of pattern recognition/handwriting recognition researchers across the globe. The major work for online handwriting in Brahmi scripts is done for smaller units as strokes and characters. So the Brahmi script's online handwriting work for larger units as words or sentences needs more attention of researchers. There are more than ten scripts Brahmi scripts. The number of alphabets (basic and compound characters) for most of Brahmi scripts is more than 250 [1]. The present work has been done for online Tamil handwriting recognition. The Tamil script is used by Tamils and Tamil speakers in India. The Tamil script is not a separated or lonely script; it has many similarities with rest scripts. As an illustration, the stroke and symbol order variation problem/challenge is common for most of Indic scripts and it is also present in Tamil script.

In direction of recognition for online handwritten strokes/characters or words in any Indic or non-Indic script, the primary or initial task is to know the basic units of writing. The basic writing unit of scripts is known as symbol. The Tamil script symbol set is presented in figure 1. In this figure, the symbols 0-10 are vowels, the symbol 11 is the special symbol, the symbols 12-33 are consonants and these have implicit vowel sound, and the symbols 72-80 represent the matras. When symbol 72 is placed over the Tamil script consonants, the consonants are changed to their half forms. The symbol 81 is always used with symbol 75. The symbols 81 and 82 are conjuncts. The symbol 83 is the dot/period. The remaining symbols are formed by using the combination of consonants and vowels, and these are called syllabic units.



அ 0	ஆ 1	இ 2	ஈ 3	உ 4	ஊ 5	எ 6	ஏ 7	ஐ 8
ஒ 9	ஓ 10	ஃ 11	க 12	ங 13	ச 14	ஞ 15	ட 16	ண 17
த 18	ந 19	ப 20	ம 21	ய 22	ர 23	ல 24	வ 25	ழ 26
ள 27	ற 28	ன 29	ஸ 30	ஷ 31	ஜ 32	ஹ 33	டி 34	டீ 35
கு 36	ஙு 37	சு 38	ஞு 39	டு 40	ணு 41	து 42	நு 43	பு 44
மு 45	யு 46	ரு 47	லு 48	வு 49	ழு 50	ளு 51	று 52	னு 53
கூ 54	ஙூ 55	சூ 56	ஞூ 57	டூ 58	ணூ 59	தூ 60	நூ 61	பூ 62
மூ 63	யூ 64	ரூ 65	லூ 66	வூ 67	ழூ 68	ளூ 69	றூ 70	னூ 71
் 72	ா 73	ி 74	ீ 75	ு 76	ூ 77	ெ 78	ே 79	ை 80
ஸ்ரீ 81	க்ஷ 82	. 83						

Figure 1. Symbol set for Tamil

All Tamil writers do not write the Tamil text in similar way and the same is present for Tamil digital handwriting and there is high degree of changeability and unpredictability of writing style of Tamil writers for writing Tamil text. Isomorphic nature of strokes or symbols is the major challenge for online Tamil handwriting recognition. Isomorphic nature means the strokes or symbols with similar shapes. There are two reasons for these types of strokes or symbols. Some Tamil strokes/symbols have similar shapes originally and others because of the writing style of writers. The stroke order and size variation, symbol order and size variation, stroke connections and stroke shapes variations, multiple characters composition in a single stroke and the presence/absence of *shirorekha* in handwritten Tamil text are the other challenges for Tamil online handwriting recognition.

This work is a vital step for online Tamil handwriting recognition and it will be proved a useful work for researchers and readers in the direction of online Tamil handwriting recognition.

LITERATURE SURVEY

For Indic scripts handwriting recognition in online mode, it is very important to study and analyze the previous work done for online handwritten Tamil text recognition in past. The existing work for Tamil script online handwritten text recognition has been done for strokes, characters and words. Thus there is availability of work for smaller units as well as larger units. The maximum existing work for online handwritten Tamil text recognition is for strokes and characters. Like other Indic scripts, the most of work for Tamil script online handwriting recognition has been done in recent two decades. We have surveyed the key work done for Tamil script online handwriting in recent decade. The table 1 presents the recognition results for Tamil online handwriting recognition in the most recent past.

Table 1.Online handwriting recognition results for Tamil script

Sr. No.	Authors and References	Year	Units	Classification Techniques	Recognition Rate (%)
1	Kunwar et. al. [2]	2014	Characters	Bayesian network (BN)	83.85
2	Kunwar et. al. [2]	2014	Characters	Random BN (RBN)	86.10
3	Kunwar et al. [2]	2014	Characters	Online RBN (ORBN)	87.80
4	Kunwar et. al. [2]	2014	Characters	Semi-supervised ORBN (SSORBN)	88.48
5	Kunwar et. al. [2]	2014	Characters	Naive bayesian (NB)	78.26
6	Kunwar et. al. [2]	2014	Characters	Support vector machine (SVM)	90.68
7	Kunwar et al. [2]	2014	Characters	Hidden Markov model (HMM)	87.82
8	Urala et. al. [3]	2014	Words	SVM + bigram	89.2 (Symbol level accuracy, GNote data), 74.5 (Word level accuracy, GNote data)
9	Urala et. al. [3]	2014	Words	SVM + bigram	83.22 (Symbol level accuracy, tablet PC data), 54.2 (Word level accuracy, tablet PC data)
10	Urala et. al. [3]	2014	Words	SVM	78.52 (Symbol level accuracy, tablet PC data), 40.05 (Word level accuracy, tablet PC data)
11	Chowdhury et. al. [4]	2013	Characters	Levenshtein distance metric	85
12	Sundaram and Ramakrishnan [5]	2013	Words	Dominant overlap criterion segmentation (DOCS), SVM	50.9
13	Sundaram and Ramakrishnan [5]	2013	Words	Attention feedback segmentation (AFS), SVM	64.9
14	Bharath and Madhvanath [6]	2012	Words	HMM	91.8
15	Sundaram and Ramakrishnan [7]	2011	Words	Dominant overlap segmentation (DOS), SVM	86.9 (Symbol level accuracy)
16	Mondal et. al. [8]	2010	Characters	Point-float feature, HMM	84.67
17	Mondal et. al. [8]	2010	Characters	Point-float feature, Multilayer perceptron (MLP)	84.98
18	Mondal et. al. [8]	2010	Characters	Chain-code feature, HMM	92.10
19	Mondal et. al. [8]	2010	Characters	Chain-code feature, MLP	91.80
20	Bharath and Madhvanath [9]	2007	Words	HMM	94.49, 93.17 and 92.15 for 5k, 10k and 20k words, respectively

a) Online handwriting recognition for Tamil script strokes and characters

A considerable work has been carried to develop standard/benchmarked databases for online handwritten isolated character recognition in Indic scripts. Most databases are also available at free of cost. In 2010, freely available datasets for Indic scripts motivated Mondal et. al. [6] to present the benchmarked recognition results for online handwritten isolated Tamil character recognition also. In their study, they have used the point-float and chain code histogram feature values to represent the feature vectors. In addition to it, they used the MLP, nearest neighbour and HMMs for classification task. They presented recognition results for six different combinations by using every classifier with every feature extraction technique. On the basis of their experimentation, they concluded that the chain code histogram feature values provide the best accuracy results without considering the type of classifier employed. They have also provided the conclusion that the best recognition results are attained by using NN classification technique irrespective the point-float and chain code histogram based feature extraction technique employed. In 2013, Chowdhury et. al. [4] employed Levenshtein distance metric to recognize isolated online handwritten Tamil characters and achieved recognition accuracy as 85%. In their work, they considered the train and test data samples' similarity for online handwritten Tamil character recognition and rest Brahmi scripts as Devanagari, Bangla, and Telugu also. In their study, to find the similarity between samples, they used Levenshtein distance metric and feature vector based on shape and position information. One year later, in 2014, Kunwar et. al. [2] employed Bayesian networks based techniques for Tamil character online handwriting recognition. This was the first ever work for Brahmi scripts where online learning of handwritten characters was employed in a semi-supervised environment. Kunwar et. al. [2] used the BN, RBN, ORBN and SSORBN for online handwritten Tamil character handwriting recognition and made conclusion that the RBN gives the best results in comparison to naive Bayes or Bayesian networks.

b) Online handwriting recognition for Tamil script words

In 2007, Bharath and Madhvanath [5] employed the HMM based word modelling for online handwritten Tamil word recognition where they used different sizes of Tamil words datasets for experimentation and attained best recognition with lexicon size of 1000 words. In their work, they recognized Tamil words in writer independent mode of handwriting. In 2011, Sundaram and Ramakrishnan [7] proposed a novel segmentation approach for Tamil words' segmentation and recognition in online mode of handwriting. The lexicon free segmentation technique proposed by them is applicable to all other Indic scripts as well as non-cursive non-Indic scripts also. In 2012, Bharath and Madhvanath [5] have taken the key work done for Latin, CJK and Arabic scripts in consideration and they referred and employed the outcome of non-Indic scripts' study in Indic scripts' online handwriting recognition for larger units (words). In Indic scripts, they have also used these studies for Tamil script. They studied and analyzed the similarities and differences of Tamil script and other Indic and non-Indic scripts for handwriting recognition in online mode. As an illustration, the problem of stroke order variation in Tamil is also faced by Latin and other Indic scripts. On the other hand, the problem of symbol order variation faced by Tamil script is not faced by Latin or CJK scripts. In 2013, Sundaram and Ramakrishnan [5] proposed a segmentation technique that uses two modules to segment online handwritten Tamil words. Their proposed segmentation technique was lexicon free and script dependent, and it solves the over and under segmentation very well. In 2014, Urala et al. [3] described a complete system to recognize isolated online handwritten Tamil words. They have also enhanced their study to recognize a paragraph of writing. In their work, they described all phases (segmentation, preprocessing, feature extraction, classification and bigram-based post processing) of online handwritten Tamil word recognition in detail.

CONCLUSION

This study has presented the most recent online handwriting recognition work done for Southern Asian/Indian script Tamil. This study has been carried out work for recognition of characters and words both. In former studies, it has been seen that the most of works for Tamil script online handwriting has been done for the smaller units. The present work is a vital step for future researchers and readers to do further/advance research work for online handwritten Tamil text recognition. This study has presented that there is the great need to carry out further research work in online Tamil handwriting recognition for larger units as words and sentences. For online handwritten Tamil sentence recognition, a great work is to be done yet.

REFERENCES

[1] B. B. Chaudhuri and U. Pal (1997). "An OCR system to read two Indian language scripts: Bangla and Devanagari (Hindi)", In: *Proceedings of 4th International Conference on Document Analysis and Recognition*.

[2] R. Kunwar, U. Pal, and M. Blumenstein (2014). "Semi-Supervised Online Bayesian Network Learner for Handwritten Characters Recognition", In*: Proceedings of 22nd International Conference on Pattern Recognition*, pp. 3104–3109.

[3] K. B. Urala, A. G. Ramakrishnan and S. Mohamed (2014). "Recognition of open vocabulary, online handwritten pages in Tamil script", *International Conference on*

Signal Processing and Communications, pp. 1–6.

[4] S. D. Chowdhury, U. Bhattacharya, and S. K. Parui (2013). “Online Handwriting Recognition Using Levenshtein Distance Metric”, In: *Proceeding of 12th International Conference on Document Analysis and Recognition*, pp. 79–83.

[5] S. Sundaram and A. G. Ramakrishnan (2013). “Attention-Feedback Based Robust Segmentation of Online Handwritten Isolated Tamil Words*”, ACM Transactions on Asian Language Information Processing,* 12(1).

[6] A. Bharath and S. Madhvanath (2012). “HMM-Based Lexicon-Driven and Lexicon-Free Word Recognition for Online Handwritten Indic Scripts”, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 34 (4), pp. 670–682.

[7] S. Sundaram and A. G. Ramakrishnan (2011). “Lexicon-free, novel segmentation of online handwritten Indic Words”, In: *Proceedings of 11th International Conference on Document Analysis and Recognition*, pp. 1175–1179.

[8] T. Mondal, U. Bhattacharya, S. K. Parui, and K. Das (2010). “On-line handwriting recognition of Indian scripts - the first benchmark”, In: *Proceeding of the 12th International Conference on Frontiers in Handwriting Recognition*, pp. 200–205.

[9] A. Bharath and S. Madhvanath (2007). “Hidden Markov Models for Online Handwritten Tamil Word Recognition”, In: *Proceedings of 9th International Conference on Document Analysis and Recognition*.

Corresponding Author

Sukhdeep Singh*

Assistant Professor in Computer Science, Mata Sahib Kaur Girls College, Talwandi Bhai, Ferozepur, Punjab, India

sukhdeep13@pu.ac.in