Lexicon-free, novel segmentation of online handwritten Indic words.

Suresh Sundaram, A G Ramakrishnan
Medical Intelligence and Language Engineering Laboratory
Department of Electrical Engineering, Indian Institute of Science,
Bangalore-560012, India
suresh,ramkiag@ee.iisc.ernet.in

***Abstract*—Research in the field of recognizing unlimited vocabulary, online handwritten Indic words is still in its infancy. Most of the focus so far has been in the area of isolated character recognition. In the context of lexicon-free recognition of words, one of the primary issues to be addressed is that of segmentation. As a preliminary attempt, this paper proposes a novel script-independent, lexicon-free method for segmenting online handwritten words to their constituent symbols. Feedback strategies, inspired from neuroscience studies, are proposed for improving the segmentation. The segmentation strategy has been tested on an exhaustive set of 10000 Tamil words collected from a large number of writers. The results show that better segmentation improves the overall recognition performance of the handwriting system.**

***Keywords*-Online Tamil symbol recognition ; Dominant overlap Segmentation (DOS) ;Feedback segmentation (FS); Stroke Group; Support Vector Machine (SVM)**

I. Introduction

Recognition of cursive handwriting has been addressed for Latin scripts in [1], [2]. Except Bangla, handwriting in Indian languages is hardly cursive. People write such that compound characters or aksharas in a word normally do not touch one another, though there can be overlaps in the horizontal direction. It is very unlikely for two or more symbols to be written in a single stroke. Except for a few selected works in [3], [4], challenges dealing with recognition of online handwritten words in Indian languages have not been adequately addressed. The published literature till date mostly focus on the issue of isolated character recognition [5], [6].

Literature deals with both segmentation-free and segmentation-based approaches for word recognition. Segmentation-free methods [7], also called holistic recognition, recognize the word as a whole using suitable features. Conversely, other techniques consider a word as a collection of segmentable symbols [8], [9]. In [10], [11], a two level segmentation scheme for Chinese symbols is reported.

This paper proposes a novel, script-independent, lexicon-free method for segmenting online handwritten Indic words to their constituent symbols. Though the proposed segmentation approach is applicable to any non-cursive Indic script, we use Tamil for illustrating the examples in this work [12]. Lexicon-free approach is useful in certain applications like form-filling, wherein, it is not feasible to invoke a finite lexicon to capture all possible proper names and addresses. To the best of our knowledge, there is no reported comprehensive work on the recognition of online handwritten words using segmentation. Our approach is motivated from studies in the area of neuroscience [14], wherein extensive feedback paths from the LGN to the cortical areas aid in either inhibiting and facilitating the responses of LGN relay cells. In this work, we employ feedbacks from features and classifier likelihoods to correctly segment a given online Tamil word into its constituent symbols. The publicly available IWFHR database [13] is used for learning various statistics about the 155 Tamil symbols. Since the focus of this work is on segmentation, 10,000 handwritten Tamil words are collected using a custom application running on a tablet PC from the students and teachers of six high schools in the South Indian state of Tamil Nadu. Hereinafter, we refer to this collection as the 'MILE database'. Statistics derived from the IWFHR database can be used to analyze the symbols in the MILE database, since our data has been acquired with the same resolution as that of the IWFHR data-set.

II. Dominant Overlap Segmentation

A handwritten Tamil word is a sequence of *n* strokes $W = \{\widetilde{s}^1, \widetilde{s}^2....., \widetilde{s}^n\}$. In general, there is significant horizontal x-overlap between the strokes of the same symbol in the case of multi-stroke Tamil characters. Initially, based on the overlap of the bounding boxes of successive strokes, the word is grossly segmented to output stroke groups. The 'Dominant Overlap Segmentation' (DOS) merges the heavily overlapping successive strokes as stroke groups, each of which is possibly a valid Tamil symbol.

If there is significant x-overlap between the k^{th} stroke group S_k and its successive stroke, the two are merged as the new S_k. Otherwise, the successive stroke begins a new stroke group S_{k+1}. In this way, all the strokes of the word are segmented. Let the minimum and maximum x-coordinates of the bounding box (BB) of the i^{th} stroke $\widetilde{s}^i$ be denoted by (x_m^i, x_M^i). Given the current stroke $\widetilde{s}^c$, its overlap O_k^c

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.237

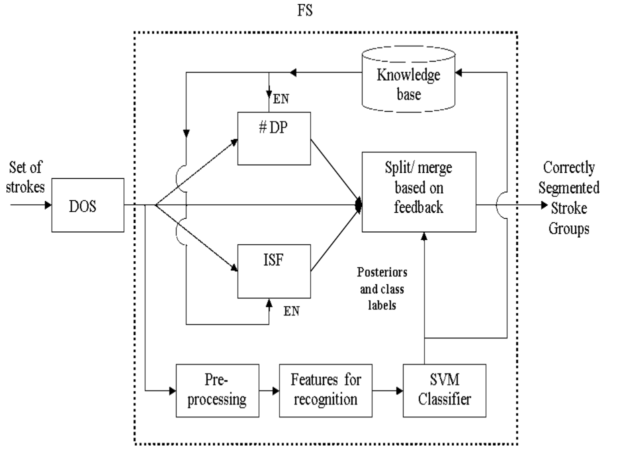
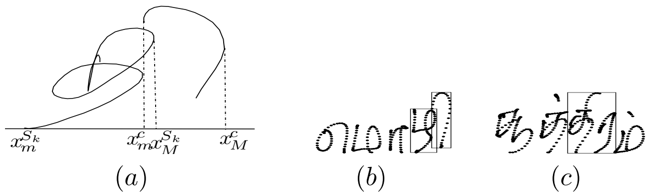


Fig. 1. Lexicon-free symbol segmentation.(DOS- Dominant Overlap Segmentation module; #DP- No. of dominant points; ISF- Inter-stroke features; EN- Enable Signal; FS- Feedback Segmentation module.)

with the previous stroke group S_k is defined as

$$O_k^c = \max\left(\frac{x_M^{S_k} - x_m^c}{x_M^{S_k} - x_m^{S_k}}, \frac{x_M^{S_k} - x_m^c}{x_M^c - x_m^c}\right) \quad (1)$$

where $x_M^{S_k}$ and $x_m^{S_k}$ are the maximum and minimum x-coordinates of the BB of the k^{th} stroke group. Fig.2(a) illustrates the parameters used to determine O_k^c. Successive strokes are merged if O_k^c exceeds a threshold T_0. The DOS step results in a set of p stroke groups, where $p <= n$. However, the DOS may output invalid patterns (Fig.2 (b) and (c)). Errors arise due to both over-segmentation and under-segmentation. The symbol ழி in the word பொழி (Fig.2(b)) is segmented into 2 stroke groups, as shown by separate BBs. The DOS outputs 5 stroke groups instead of 4. In Fig.2(c), the symbols தி and ர of the word சூத்திரம் merge to a single stroke group, which is highlighted by a single BB. In this case, DOS outputs 4 stroke groups instead of 5. Such segmentation errors reduce the recognition rate of the handwriting system.



(a) (b) (c)

Fig. 2. DOS. (a) Computing the degree of overlap O_k^c between two contiguous strokes. (b)and (c) Incorrect segmentation by DOS for threshold $T_0 = 0.2$.

III. Feedback Segmentation

Feedback segmentation (FS) detects possible errors in the generated stroke groups from the DOS, and refines the segmentation by merge, retain or split operations to output valid stroke groups or symbols. For the examples shown in Fig.2 (a) and (b), if the FS module is successful, it should output 4 and 5 stroke groups, respectively. The stroke groups output by the DOS are regarded as tentative candidates for valid Tamil symbols. Feedback segmentation may modify the number of stroke groups output by DOS, based on specific criteria proposed in this work. The ensuing stroke groups are dealt with as valid symbols for the given word. The blocks (shown under dotted rectangle in Fig.1) are employed in the feedback segmentation module and are discussed in the following sections.

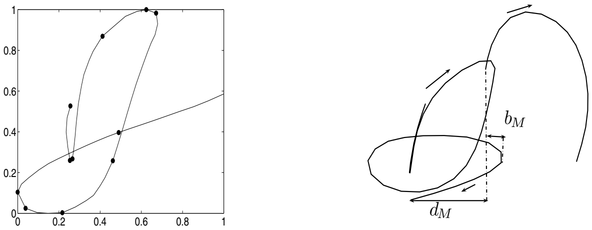
IV. Creation of Knowledge Base

A set of features is derived for each stroke group obtained from the DOS step. The statistics obtained for each of these features from the symbols in the IWFHR dataset have been effective in detecting and correcting possible segmentation errors.

1) **Number of Dominant Points**: The number of dominant points (# of DPs) of a stroke group is used as a cue in this work. Given a preprocessed stroke group, we begin by marking the first pen position as a DP. Starting from the current DP, the absolute value of the angle between pen directions at successive points is computed and accumulated along the online trace. The accumulation step is done as long as the cumulative sum T_s is less than a threshold T_θ. The pen position, at which $T_s \geq T_\theta$, is marked as the next dominant point and the process continues till the end of the trace. The resulting # of DPs extracted is used as a feature descriptor. Fig.3 (a) presents the DPs for the stroke group ச with T_θ set to 45^o. Pre-processing a stroke group comprises the steps discussed in [5]. For the k^{th} stroke group, we denote the number of dominant points by N^{S_k}.
2) **Inter-stroke features (ISF)** : These features apply to stroke groups comprising m strokes ($m > 1$)
   a) The horizontal distance b_i from the *BB* *x*-maxima of the i^{th} stroke to the first point of the $(i+1)^{th}$ stroke is recorded. The maximum of the computed widths (denoted by b_M) is used as a feature.
   b) The horizontal gap d_i between last point of the i^{th} stroke to the first point of the $(i+1)^{th}$ stroke is computed. The maximum of the computed widths (denoted by d_M) is used as a feature.

The ISF is illustrated for the stroke group கி (Fig.3 (b)), written in 2 strokes. As will be discussed in the following sections, the ISF detects possible merges of valid symbols in a stroke group, while # of DPs identify probable broken symbols.

By employing the features described, a knowledge base is created , wherein for each symbol ω_c in the IWFHR dataset, the following statistics are generated.



(a) (b)

Fig. 3. Extraction of (a) dominant points for a character. (b) inter-stroke features for a stroke group. Here, $d_M > 0$, $b_M < 0$. The direction of the trace is shown by arrows.

Table I
FREQUENCY DISTRIBUTION OF DOMINANT POINTS IN WRONGLY-SPLIT SYMBOLS OF THE IWFHR DATASET.

# DP	Frequency
1-3	45
4-6	187
7-10	170
11-13	41
13-15	11

1) Maximum number of dominant points $N_M^{\omega_c}$ over all valid samples.
2) Maximum horizontal inter stroke gap $d_{thr}(\omega_c)$ over all valid samples.

V. Detect and merge over-segmentation

The DOS is applied to each of the training samples of symbols in the IWFHR dataset. The presence of two or more stroke groups in a given sample indicate an over-segmentation error. The frequency distribution of the # of DPs for the symbols (in the IWFHR dataset) broken by DOS is presented in Table.I. A stroke group possessing less than 16 DPs may correspond to a part of a symbol that has been over segmented. Let S_k correspond to a stroke group that is likely to be a broken symbol. Consider $S_{N(k)}$ to be the neighboring stroke group whose BB is closest to that of S_k. The x-y coordinates of the trace of S_k and $S_{N(k)}$ are independently pre-processed and sent for recognition to generate the likelihoods for the most probable symbols $P(\omega_{top}^k)$ and $P(\omega_{top}^{N(k)})$. The SVM is used as the classifier.

1) The stroke groups are merged whenever, $P(\omega_{top}^k) < P_{thr}(\omega_{top}^k)$ or $P(\omega_{top}^{N(k)}) < P_{thr}(\omega_{top}^{N(k)})$. $P_{thr}(\omega_{top}^k)$ and $P_{thr}(\omega_{top}^{N(k)})$ represent the minimum likelihood returned by the SVM across all correctly recognized samples of ω_{top}^k and $\omega_{top}^{N(k)}$ in the IWFHR Test set.
2) Let S_M represent the stroke group, obtained by merging S_k with $S_{N(k)}$. For a possible merge, we require the average likelihood of ω_{top}^k and $\omega_{top}^{N(k)}$ to be less than $P(\omega_{top}^M)$ for S_M. However, for avoiding any unintentional merges, we ensure that the maximum horizontal inter stroke gap in S_M is less than the maximum possible horizontal gap $d_{thr}(\omega_{top}^M)$. We obtain the value of $d_{thr}(\omega_{top}^M)$ from the knowledge base.

Fig. 4 present an illustration of a word தொண்டர், wherein the symbol ண் suspected to be broken in the DOS gets corrected in the FS module.

VI. Detect and split under-segmentation

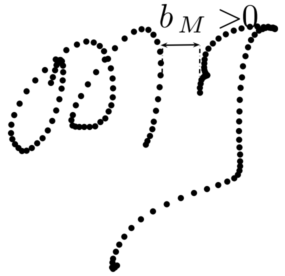


Fig. 5. Distinct symbols wrongly merged by DOS. Here $b_M > 0$.

A detailed analysis of raw stroke groups (comprising multiple strokes) obtained from applying the DOS step on the words from the MILE database reveal that stroke groups satisfying $b_M > 0$ may correspond to valid symbols that have been merged, resulting in an under-segmentation error. Referring to Fig.5, the 2 symbols ை and ர are merged to a stroke group. Being an outlier, the SVM in general provides a low likelihood to the stroke group ை ர.

Let b_M correspond to the inter stroke gap between q^{th} and $(q+1)^{th}$ strokes in S_k respectively. Accordingly, the two valid symbols s^1 and s^2 merged in S_k can be written as $s^1 = \{S_k^1, S_k^2, ...S_k^q\}$ and $s^2 = \{S_k^{q+1}, S_k^{q+2}, ...S_k^m\}$. The x-y coordinates of s^1 and s^2 are in turn pre-processed and subsequently recognized to generate confidence likelihoods $P_j^* = \max_i \quad P(\omega_i|\mathbf{x}^{s^j}) \qquad j = 1, 2$. We favor splitting the stroke group S_k into s^1 and s^2 whenever $\frac{\sum P_j^*}{2} \geq P(\omega_{top}|\mathbf{x}^{S_k})$. Here ω_{top} represents the most probable symbol returned by the SVM for S_k. For the scenario, where the inequality is not satisfied, stored information from the knowledge base is employed to split S_k.

1) If $N^{S_k} \geq N_M^{\omega_{top}}$, the signal EN is enabled (refer Fig.1) and we proceed ahead in segmenting S_k into 2 valid symbols s^1 and s^2.
2) If $d_M \geq d_{thr}(\omega_{top})$, EN is enabled and we segment S_k. d_M is computed as described in Sec.IV.

Fig.6 illustrates the case wherein the erroneous stroke group வீர at the start of the word வீராங்கனை is segmented correctly to 2 valid symbols வீ and ர respectively.

VII. Performance on the MILE database

Before applying the proposed feedback segmentation technique on the Tamil words, the parameters of SVM are tuned by testing its performance on the IWFHR Competition

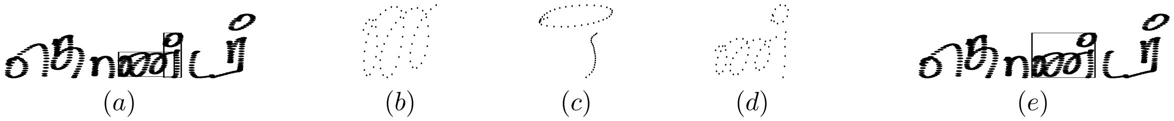


Fig. 4. Correction of over-segmentation. (a) Fourth symbol is split into two by DOS. (b) This Stroke group has a low posterior probability and is suspected to be a part of a symbol. (c) The second split part of the symbol also has low posterior probability. (d) Merged symbol has higher likelihood. (e) Correctly segmented word.

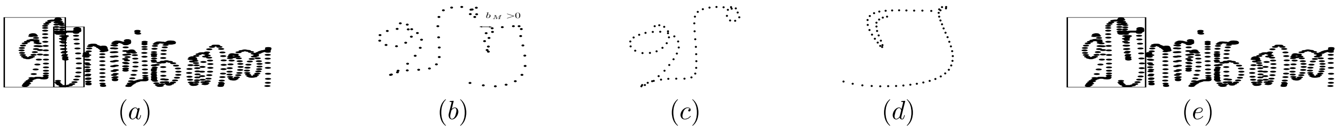


Fig. 6. Correction of under-segmentation. (a) A word under-segmented by DOS (b) Stroke group with $b_M > 0$ and low likelihood is suspected to comprise two valid symbols. (c) and (d) the individual split symbols, when recognized, have higher average likelihood. (e) Word correctly segmented by FS.

Table II

MERGER OF TWO SYMBOLS BY DOS, SPLIT BY FS AND CONSEQUENT IMPROVEMENT IN RECOGNITION. THE INCORRECTLY RECOGNIZED SYMBOLS ARE SHOWN WITHIN BOXES.

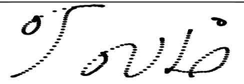
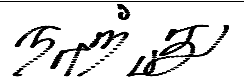
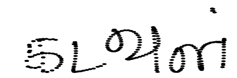
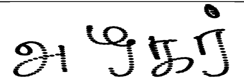
Input Word	Recognized symbol after DOS	Recognized symbol after FS
	தும்	ஏலம்
	நசஷ்நு	நாற்பது

Table III

SPLITTING OF SYMBOLS INTO TWO STROKE GROUPS BY DOS, CORRECT SEGMENTATION BY FS AND CONSEQUENT IMPROVEMENT IN RECOGNITION. THE INCORRECTLY RECOGNIZED SYMBOLS ARE SHOWN WITHIN BOXES.

Input Word	Recognized symbol after DOS	Recognized symbol after FS
	கடவுளசூ	கடவுள்
	சஹூழகர்	அழகர்

Test set. The x and y coordinates of the pre-processed symbols are used as features. A recognition performance of 86% is achieved on the test set with the RBF kernel parameters $C = 5$ and $\gamma = 0.2$.

We now demonstrate the impact of the proposed FS strategies on the MILE database. A few sample words correctly segmented by the algorithm are shown in Tables II and III. DOS on each word in Table II led to an under-segmentation error. On the other hand, one valid symbol in each word in Table III is wrongly split. The incorrect segmentation by the DOS in turn increases the symbol recognition errors (as indicated by the squared boxes in the second column of the tables). However, all the constituent symbols of these words are recognized correctly after the FS step (as observed from the third columns). Across the 10000 words in the MILE database comprising 53026 symbols, a segmentation rate of 99.7% is achieved at symbol level after FS. The segmentation with feedback in turn improves the symbol recognition rate from 83% (with DOS alone) to 86.9%.

We finally address the drawbacks of the proposed algorithm. Segmentation fails in scenarios where symbols are written as a different temporal sequence not frequently encountered in practice. Moreover, the methods are not robust in merging symbols comprising large horizontal inter-stroke gaps, that are comparable to the horizontal inter-character gaps. Given that there is no prior work done in segmenting online Indic words, it is difficult to compare our method to a benchmark. However, the features being script independent, there is scope in adopting similar feedback based methodologies to segment words in other Indic scripts such as Kannada, Telugu and Malayalam.

VIII. Conclusion

In this work, we present a novel unlimited vocabulary, lexicon free segmentation approach for online Indic words. The given word is segmented in the DOS into a set of stroke groups. Using dominant point and inter stroke features, segmentation errors, if any, are detected. The stroke groups suspected to be erroneous are corrected with feedback strategies to form valid symbols (FS module). The reduction of the segmentation errors in the FS module in turn leads to an improvement in the performance of the handwriting recognition system. The high success rate of segmentation holds promise in recognition of online handwritten words such as proper names and addresses, where it is not possible to invoke a finite lexicon.

Acknowledgements

The authors would specially like to thank TDIL, DIT for funding this research work. The support from Prof. Arun Sripathi, Centre for Neuroscience, IISc and Vikram Lakavalli is also acknowledged.

References

[1] F Camastra, A SVM-based cursive character recognizer, PR(40), pp.3721-3727,2007.

[2] A W Senior, A J Robinson, An Off-Line Cursive Handwriting Recognition System, IEEE Trans PAMI(20), pp.309-321, 1998.

[3] U Bhattacharya, A Nigam, Y S Rawat, S K Parui, An Analytic Scheme for Online Handwritten Bangla Cursive Word Recognition. Proc.ICFHR, pp.320-325, 2008.

[4] A Bharath, S Madhvanath, Hidden Markov Models for Online Handwritten Tamil Word Recognition, Proc.ICDAR, pp. 506-510, 2007.

[5] N Joshi, G Sita, A G Ramakrishnan, S Madhavanath, Comparison of Elastic Matching Algorithms for Online Tamil Handwritten Character Recognition, Proc. IWFHR, pp.444-449,2004.

[6] U Bhattacharya, B K Gupta, S Parui, Direction Code Based Features for Recognition of Online Handwritten Characters of Bangla, Proc.ICDAR(1), pp.58-62,2007.

[7] S Madhvanath, V Govindaraju, The Role of Holistic Paradigms in Handwritten Word Recognition, IEEE Trans. PAMI.23(2),pp.149-164, 2001.

[8] M Nagakawa, B Zhu, M Onuma, A model of online handwritten Japanese text recognition free from line direction and writing format constraints, IECIE Trans on Info. and Sys, pp.1815-1822, 2005.

[9] C L Liu, H Sako, H Fujisawa, Effects of Classifier Structures and Training Regimes on Integrated Segmentation and Recognition of Handwritten Numeral Strings, IEEE Trans.PAMI, 26(11), pp.1395-1407, 2004.

[10] X Gao, P M Lallican, C Viard-Gaudin, A Two-stage Online Handwritten Chinese Character Segmentation Algorithm Based on Dynamic Programming, Proc.ICDAR, pp.735-739, 2005.

[11] S Y Zhao ,Z R Chi ,P F Shi, Two-stage segmentation of unconstrained handwritten Chinese characters. PR (36), pp.145-156, 2003.

[12] Suresh Sundaram, A G Ramakrishnan, Verification based Segmentation approach for online words. Indian Patent Office Reference. No: 03974/CHE/2010

[13] www.hpl.hp.com/india/research/penhw-interfaces-1linguistics.html

[14] A M Sillito, H E Jones, Corticothalamic interactions in the transfer of visual information, Philos Trans R Soc Lond B Biol Sci, pp.1739-1752, 2002.