உத்தமம்
INFITT
உலகத் தமிழ்த் தகவல் தொழில்நுட்ப மன்றம்
International Forum for Information Technology in Tamil

Penn
University of Pennsylvania

மாநாட்டுக் கட்டுரைகள்
CONFERENCE PAPERS

TAMIL INTERNET 2011

தமிழ் இணையம் 2011

Neuroscience inspired segmentation of handwritten words

A G Ramakrishnan and Suresh Sundaram
MILE Lab, Dept of Electrical Engineering, Indian Institute of Science, Bangalore.

The challenge of segmenting online handwritten Tamil words has hardly been investigated. In this paper, we report a neuroscience-inspired, lexicon-free approach to segment Tamil words into its constituent symbols (recognizable entities). Based on a simple dominant overlap criterion, the word is grossly segmented into candidate symbols (stroke groups). However, this segmentation is not fully reliable because of varying writing styles resulting in varying levels of overlap. Taking cues from vertebrate visual perception, we utilize both feature based attention and feedback from the classifier to detect possible wrong segmentations. This attention-feedback segmentation (AFS) strategy splits or merges the stroke groups to correct the segmentation errors and forms valid symbols. This maiden attempt on segmentation is tested on 10000 handwritten words collected from hundreds of writers. The efficacy of AFS in segmentation and improving the recognition performance of the handwriting system is amply demonstrated. Our results show a segmentation accuracy of over 99% at symbol level.

Need for segmenting handwritten words

Since attempts to segment cursively handwritten English words have largely failed, researchers working on Indic scripts too feel that it is not advisable to try to segment individual characters from handwritten documents. However, we firmly believe that it is not only possible, but also something that ought to be done, if one is interested in recognizing words such as proper names appearing in the name and address fields of handwritten forms. Thus, this opens up the possibility of developing a recognizer that can handle unrestricted vocabulary, including any unusual word of foreign origin, such as names of people or places from other countries. Thus, we believe that our work is the first of its kind in proposing an approach for handwriting recognition that does not limit the writer from writing any text of any origin.

Motivation for Attention-Feedback Segmentation

Traditional pattern recognition [1, 3-5, 8, 10, 13-15] primarily follows a feedforward architecture, whereas the same in mammalian brain involves complex feedback structures. Studies on visual perception in primates demonstrate the effect of attention on the response of the visual neurons [2]. Feature based attention biases the neuronal responses as though the attended stimulus was presented alone. Also, shifting spatial attention from outside to the inside of the receptive field increases the neuronal responses. Motivated by these observations, we incorporate local feature based attention to correct and improve segmentation [9]. Further, studies on visual pathways show extensive feedback from the cortex to the lateral geniculate nucleus (LGN), which have both inhibitory and facilitatory effects on the responses of LGN relay cells. In our work, we use feedback based on features as well as from the classifier posterior probabilities to rectify any incorrect segmentation by regrouping the strokes. Thus, we call our approach as 'attention-feedback' strategy for segmentation.

Further, studies on scene perception by humans [6] indicate that visual processing follows a top-down approach. The global cues characteristing the visual object, that appear within the visual span, are perceived before the local features. The human perceptual system treats every scene as if it were in the process of being focussed or zoomed in on, whereas initially, it is relatively less distinct. Moreover, the human perceptual processor has the capability to select parts of the input stimulus that are worth to be paid attention to. Motivated with these observations from the field of neuroscience, we present a segmentation strategy that first works on the global feature of overlap to output candidate Tamil stroke groups for the given input strokes. By analyzing local features characteristic to the given input pattern, we reevaluate the segmentation and modify the segmentation when found necessary. The localized features are derived by zooming on paying attention to specific parts of the online trace. Essentially, we adopt a multi-pass system, wherein fine grained processing is guided by the prior cursory (global) processing.

Data used for the study

The 155 distinct Tamil symbols (comprising 11 vowels, 23 base consonants, 23 pure consonants, 92 CV combinations and 6 additional symbols) are presented in Appendix A. The publicly available corpus of isolated Tamil symbols (IWFHR database) is used for learning various statistics about Tamil symbols. The primary focus of this work is to address the challenges of segmentation. Towards this purpose, Tamil words are collected using a custom application running on a tablet PC and saved using a XML standard [7]. High school students from across 6 educational institutions in Tamil Nadu contributed in building the word data-base of 100, 000 words, referred to as the 'MILE Word Database' in this work [12]. Out of these, 10,000 words are used for this study. The words have been divided into 40 sets, each comprising 250 words. Owing to the comparable resolution of our input device to that used in the IWFHR dataset, statistical analysis performed on the symbols in the IWFHR database are applicable to the Tamil symbols in the MILE word database.

Dominant Overlap Criterion Segmentation

An online word can be represented as a sequence of n strokes $W = \{s_1, s_2, \ldots, s_n\}$. In the case of multi-stroke Tamil symbols, strokes of the same symbol may significantly overlap in the horizontal direction. The word is first grossly segmented based on a bounding box overlap criterion, generating a set of stroke groups. In this 'Dominant Overlap Criterion Segmentation' (DOCS), the heavily overlapped strokes are merged. A stroke group is defined as a set of consecutive strokes merged by the DOCS step, which is possibly a valid Tamil symbol.

For the k-th stroke group S_k under consideration, its successive stroke is taken and checked for possible overlap. Significant overlap necessitates the successive stroke to be merged with the stroke group S_k. Otherwise, the successive stroke is considered to begin a new stroke group S_{k+1}. The algorithm proceeds till all the strokes of the word are exhausted.

Neuroscience-inspired segmentation

The stroke groups obtained from the above dominant overlap criterion segmentation are preprocessed by smoothing, normalization and resampling into standard number of equi-arc length spaced points.

The x and y coordinates of these processed stroke groups and their first and second derivatives are used as features for recognition using a support vector machine (SVM) classifier that outputs class labels and their posterior probabilities. Obviously, DOCS being simple, does not always result in correct segmentation. Sometimes it results in over segmentation of a single multi-stroke character into two stroke groups; other times, two distinct characters get combined into a single stroke group, due to the way they are written.

Attention Features

Figure 1 shows the complete block schematic of the proposed segmentation scheme. An over-segmented symbol is usually small and hence results in low aspect ratio as well as has very few dominant points (points where the curvature is high). By paying attention to these features extracted from the stroke groups output by DOCS block, one can suspect wrong segmentation. Further, the symbols that result from over- or under-segmentation are classes that the classifier has not come across. Thus, these symbols usually result in a low confidence level of the classifier. Thus, the posterior probability of the classifier, when fed back to the input stages, can be used to invoke the computation of the attention features. The feedback, together with the attention features suggest possible resegmentation of the input strokes, resulting in new possible stroke groups. These modified stroke groups based on merger or splitting of original stroke groups, are once again recognized by the classifier after preprocessing and extraction of recognition features. An improved posterior probability of the new stroke group confirms right segmentation. Thus, the refinement in segmentation is caused based on memory, attention and feedback mechanisms prevalent in human perception. We call this as "attention-feedback segmentation (AFS)".

Commonly found segmentation issues

The two Tamil characters that ought to have a minimum of three strokes are the long /i/ (nedil) and the aydam. Since in both of these cases, in general there is no overlap between the final dot and the rest of the character, they always are over segmented into two or more stroke groups.

Pure consonants (*mey ezhuthu*), when they are written with the dot (*pulli*) beyond the base consonant, result in over segmentation too.

Characters such as /ka/, /nga/ and /ra/, which start with an initial vertical segment, are written by many with multiple strokes, with the first stroke being a simple down-going vertical line. These characters have a potential to be over segmented, if the following part of the character does not clearly overlap with the vertical line.

All CV combinations of /i/ and /I/ and the CV combinations of /u/ and /U/ with borrowed consonants such as /ja/ and /sha/ also have a tendency to be over-segmented, if the vowel matra is written with no horizontal overlap with the consonant.

Under segmentation occurs if the ending part (usually bottom extensions of /ta/ or /Ra/) of the following character goes far left below the previous character, causing significant horizontal overlap between them. At other times, people write two successive characters so closely, that there is significant overlap between them.

Naturally, in all the above cases, the simple segmentation (DOCS) is likely to result in wrong segmentation leading to erroneous recognition results.

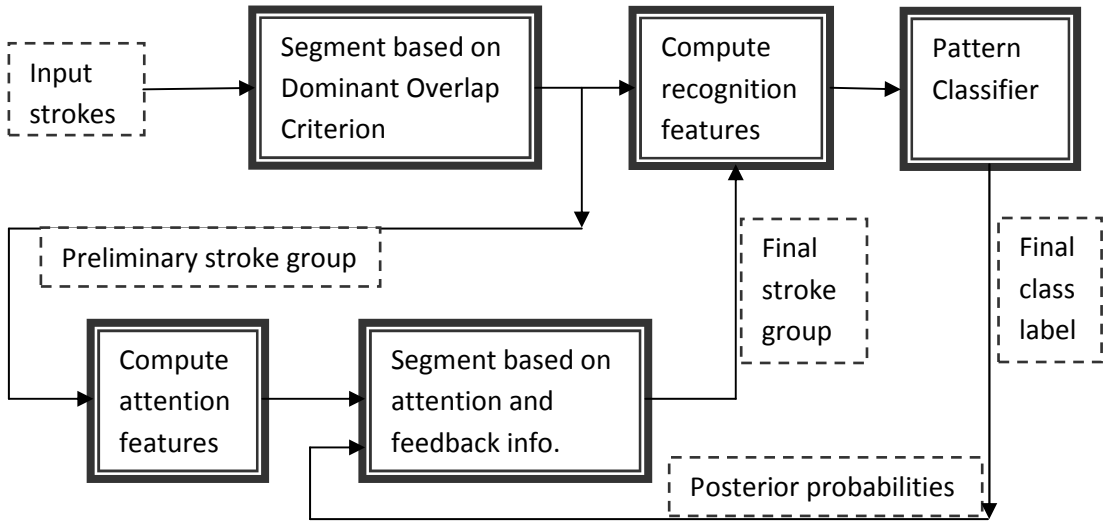

Fig. 1. Block diagram of neuroscience inspired segmentation of Tamil handwritten word [9].

Segmentation results on the MILE Tamil Word Database

The proposed techniques are tested on the subset of 10,000 words. However, to start with, we evaluate the performance on a set of 250 words (denoted as DB1), that has a significant number of errors resulting from the DOCS. Of the 103 errors, 89 (or 86%) correspond to the merging of valid symbols, and the rest, to broken symbols. The AFS module aids in properly detecting and correcting 91 (or 90%) of these errors. In addition, the methods proposed effectively merge 11 (or 78%) of the over-segmented stroke groups to valid symbols. The improvement in character segmentation rate in turn reduces the number of wrongly segmented words. It is observed that only 7 of the total 250 words remain wrongly segmented after the AFS scheme, as against 67 words after the DOCS scheme. On evaluating the performance across the database of 10000 words, we obtain a 86% reduction in character segmentation errors.

Recognition results on the MILE Database

We report experimental results demonstrating the impact of the proposed AFS strategy on the recognition of symbols in the MILE word database. Since a significant percentage of DOCS errors are corrected by AFS, a drastic improvement of 16% (from 70.5 % to 87.1 %) in symbol recognition is observed. In computing the symbol recognition rate, apart from the substitution errors, we take into account the insertion and deletion errors, caused by over-segmentation and under-segmentation, respectively. The edit distance is used for matching the recognized symbols with the ground truth data. Moreover, 11.6 % of the words, (29 additional words) wrongly recognized after DOCS, have

been corrected by the proposed technique. Across the 10000 words in the MILE Word database, an improvement of 4% (from 83 to 87%) in symbol recognition rate has been obtained.

Conclusion

In this paper, we present a maiden attempt based on significant feedback from the classifier to the input blocks such as feature extraction and segmentation, as well as the use of memory (prior knowledge) to result in a very effective segmentation of online handwritten Tamil words. This approach being general, can be extended to any other Dravidian script, as well as any other script where cursive writing is not practiced. To our knowledge, there is no reported systematic research work on segmenting the individual characters or recognizable standard symbols from online handwritten words for any Indic language that does not have the *shiro rekha* (head line). Thus, we are unable to compare the performance of our work with any other technique. However, the results are promising and have also led to improved recognition of the handwritten words [16], thus confirming the possibility of proper segmentation of online Tamil words. We intend to extend this work very soon to online Kannada handwritten words.

Acknowledgment

We thank Ms. Nethra Nayak, Mr. Rituraj Kunwar, Ms. Archana C P, Mr. Shashi Kiran, Ms. Chandrakala, Mrs. Shanthi Devaraj, Ms. Saranya and Ms. Sountheriya for their efforts in data collection and annotation, which made these experiments possible. We thank Dr. Arun Sripathi of the Centre for Neuroscience, IISc for the very useful discussions we had on visual perception. Special thanks to Technology Development for Indian Languages (TDIL), Department of Information Technology (DIT), Government of India for funding this research, as part of a research consortium on online handwriting recognition in several Indic scripts. We thank Prof. Deivasundram (University of Madras), AVM Matriculation Higher Secondary School Virugambakkam, Chennai, Govt. Boys Higher Secondary School, Sulur, Presidency College, Triplicane, Chennai and IIT Madras for contributing to the data set. We also thank Dr. Anoop Namboodiri for giving us the word level annotation tool. We thank CDAC, Pune for being a partner in finalizing the XML standard for handwritten data collection in Indic languages.

References

- N Joshi, G Sita, A G Ramakrishnan, S Madhavanath, "Comparison of elastic matching algorithms for online Tamil handwritten character recognition", Proc. IWFHR (2004) 444-449.
- G M Boynton, "Attention and visual perception. Current Opinion in Neurobiology", (15) (2005) 465-469.
- H Swethalakshmi, C Chandra Sekhar, V S Chakravarthy, "Spatiostructural features for recognition of online handwritten characters in Devanagari and Tamil scripts", ICANN (2) (2007) 230-239.
- A Bharath, S Madhvanath, "Hidden markov models for online handwritten Tamil word recognition", Proc. ICDAR (2007) 506-510.
- Amrik Sen, G. Ananthakrishnan, Suresh Sundaram, A. G. Ramakrishnan, "Dynamic space warping of strokes for recognition of online handwritten characters. IJPRAI (2009) 23(5): 925-943.

- Arun P Sripati and Carl R Olson, "Representing the forest before the trees: a global advantage effect in monkey inferotemporal cortex", The Journal of Neuroscience, June 17, 2009, 29(24):7788 –7796.
- Swapnil Belhe, Srinivasa Chakravarthy, A. G. Ramakrishnan, XML standard for Indic online handwritten database, ACM - Proceedings of the International Workshop on Multilingual OCR, 2009.
- M. Mahadeva Prasad, M. Sukumar, A. G. Ramakrishnan, Divide and conquer technique in online handwritten Kannada character recognition, ACM - Proc International Workshop on Multilingual OCR, 2009.
- Suresh Sundaram and A G Ramakrishnan, "Verification based segmentation approach for online words", Indian Patent Office Ref. No. 03974/CHE/2010.
- Shashi Kiran, Kolli Sai Prasada, Rituraj Kunwar, A. G. Ramakrishnan, "Comparison of HMM and SDTW for Tamil handwritten character recognition", Proc. 2010 IEEE International Conf Signal Processing & Communication.
- Rituraj Kunwar, Mohan P., Shashi Kiran, A. G. Ramakrishnan, "Unrestricted Kannada online handwritten akshara recognition using SDTW, Proc. 2010 IEEE International Conf Signal Processing & Communication.
- B Nethravathi, C P Archana, K Shashikiran, A G Ramakrishnan, V Kumar, "Creation of a huge annotated database for Tamil and Kannada OHR", Proc. IWFHR (2010) 415-420.
- M. Mahadeva Prasad, M. Sukumar, A. G. Ramakrishnan, "Orthogonal LDA in PCA transformed subspace", Proc. 12th International Conf Frontiers in Handwriting Recognition (ICFHR 2010), Nov 2010.
- Venkatesh N, A G Ramakrishnan, "Choice of classifiers in hierarchical recognition of online handwritten Kannada and Tamil aksharas", Jl. Universal Computer Science, 2011, Vol. 17, No. 1, pp. 94-106.
- Rakesh R, A G Ramakrishnan, "Fusion of complementary online and offline strategies for recognition of handwritten Kannada characters", Journal of Universal Computer Science, 2011, Vol. 17 (1), pp. 81-93.
- Suresh Sundaram and A G Ramakrishnan, "Attention-feedback based robust segmentation of online Tamil words", under review, Pattern Recognition, 2011.