

D.2 Project for intern: DNN based OHWR system

A Motivation

This era of technology has witnessed a great deal of efforts to make computers more “natural” in their interface. In this regard, handwriting is considered as the best option, next only to speech. Although there are a good number of English and Chinese online handwriting recognition (OHWR) systems, there are only a few for Indic languages. The OHWR system for Tamil developed by MILE Laboratory is one such system.

The MILE Tamil OHWR system is unique in the sense that it makes use of linguistic knowledge in many stages, from segmentation to post-processing. This is one of the major reasons for its high recognition rate. In addition, the knowledge of how human brain processes handwritten text was used in the development, making the engine a “brain-inspired” OHWR system. The patent pending “attention-feedback” technology is one such technology used in the system.

Although the present engine can process even paragraph level mixed data (data containing both Tamil characters and Indo-Arabic numerals) with appreciable recognition rate, it is based on kernel SVM. We are looking for committed interns to convert the system to the state-of-the-art DNN based OHWR system. Since enough data to train a deep network is not available, techniques such as transfer learning, data augmentation etc should be adopted, making this project more challenging. We believe that DNNs will help to make the system better in terms of preprocessing speed and recognition rate.

Potential Applications

- **Form filling:** Conversion of handwritten forms in Tamil to digital format for easy processing.
- **Assistive technology for the disabled:** Vocally disabled persons can write in Tamil on their mobile phone and the device will convert the handwritten text to speech.

Challenges

- Training of deep networks using limited data.
- More number of classes than Latin scripts.

B Problem statement

- To develop a DNN based Tamil OHWR system which can process paragraph level mixed data with high accuracy and low latency.

C Facilities available

- High performance cluster computing server for deep learning applications.
- Tablet PCs for testing and further data collection, if required.

D Skills required

1. Essential

Background in signal processing and machine learning

Programming skills: C++ and Python

2. Desirable

Knowledge of Keras, PyTorch etc.

Background in Natural Language Processing (NLP).

Knowledge of Tamil.

E References

1. Ramakrishnan, A. G. "The magic of automated recognition of handwriting." *Current Science* 107.2 (2014): 159-160.
2. Sundaram, Suresh, and A. G. Ramakrishnan. "Bigram language models and reevaluation strategy for improved recognition of online handwritten Tamil words." *ACM Transactions on Asian and Low-Resource Language Information Processing* 14.2 (2015): 8.
3. Urala, K. Bhargava, A. G. Ramakrishnan, and Sahil Mohamed. "Recognition of open vocabulary, online handwritten pages in Tamil script." *Signal Processing and Communications (SPCOM), 2014 International Conference on*. IEEE, 2014.
4. Sundaram, Suresh, and A. G. Ramakrishnan. "Language models for online handwritten Tamil word recognition." *Proceeding of the workshop on Document Analysis and Recognition*. ACM, 2012.