Research Journal of Computer Systems and Engineering

ISSN: 2230-8571, 2230-8563

Volume 03 Issue 01 - 2022 (January to June)

Page 35:42

Handwritten Tamil Word Pre-Processing and Segmentation Based on NLP Using Deep Learning Techniques

Dr.R.Kishore Kanna
Computer Science & Communication Engineering;Signal Processing;Bio-Engineering and Technology
Vels Institute of Science Technology and Advanced Studies
ORCHID ID-0000-0002-8004-1501
kishorekanna007@gmail.com

Iskandar Muda
Management
Universitas Sumatera Utara
ORCHID ID-0000-0001-6478-9934
iskandar1@usu.ac.id

Dr.S.Ramachandran
Mechanical and Electrical Engineering
Department of Electrical and Electronics Engineering, Paavai Engineering college
https://orcid.org/0000-0002-1579-4742
contacttoramachandran@gmail.com

Article History

Received: 22 January 2022
Revised: 14 April 2022
Accepted: 19 May 2022

Abstract

Tamil is a traditional Indian language spoken mostly among South Indians, SriLankans, as well as Malaysians. This paper proposed the novel techniques based on pre-processing and segmentation of handwritten Tamil words through NLP using threshold value based RGB image conversion to grayscale image. Then to segment this image based on line boundary detection with Alex Net based Convolutional neural network (Alex Net- CNN) in deep learning architecture. Every text is scaled in to needed pixel in the suggested system, that is then exposed to be trained. – i.e., every scaled word contains a set pixel count, which are used to train networks. The findings reveal that proposed method achieved better detection accuracy in written vocabulary knowledge that are equivalent to features extraction techniques. For numerous pictures, a descriptive analysis was performed in terms of effectiveness, accuracy, recollect, and F1 measure.

Keywords: Tamil language, Image Processing, pre-processing, segmentation, Alex Net- CNN.

1. Introduction:

The majority of higher level natural language processing tasks, including part-of-speech tagging (POS), parsing, and machine translation, start with word segmentation. You could think of it as the

challenge of accurately detecting word formations from a character string. Word segmentation can be quite difficult, especially for languages like Chinese, Japanese, and Vietnamese that lack clear word boundary delimiters [1]. Depending on how a word is defined, several word segmentation standards exist [2]. In practically every language in the world, Natural Language Processing (NLP) is a major topic of research. Computers are taught in NLP so they can comprehend and manipulate human language text and speech with ease. The goal of NLP research is to develop an understanding of how people use and comprehend natural language. To enable computer systems to understand and use natural languages and complete the needed tasks, they use appropriate tools and procedures that are capable of becoming technologically upgraded. The foundations of NLP are found in a variety of fields, including mathematics, psychology, linguistics, artificial intelligence (AI), electronic and electrical engineering, and information and computer sciences [3].

2. Related works:

Lately, an expanding pattern has been noticed for written by hand character and word acknowledgment utilizing deep learning. In [4], the model vitally examined was changed a little by applying a few adjustments, for example, the utilization of a CNN-RNN model. A significant approach CNNLSTM was proposed in [5] by utilizing CNN model alongside 1D LSTM model. One significant downside in the current models was the powerlessness to interpret inconsistent person strings independent of word reference size. To manage such situation, a Word pillar search based component is used in [6]. A LSTM-based model is additionally taken advantage of in [7] adaptable internet based manually written text acknowledgment framework. In [8], a concise outline of different approaches was completed for investigation of content based picture recovery in packed area on the MPEG dataset. A few analysts have revealed exactness as great as 98% or almost 100% for manually written digit acknowledgment [9]. A group model has been planned utilizing a blend of numerous CNN models. The acknowledgment test was completed for MNIST digits, and an exactness of 99.73% was accounted for [10].

3. System model:

This section discuss about the proposed architecture of Tamil word segmentation. The input has been pre-processed for image threshold based RGB image to grayscale conversion and segment the image of handwritten Tamil word line by line through line boundary detection with AlexNet-CNN. The architecture of proposed technique is given in fig.,-1.

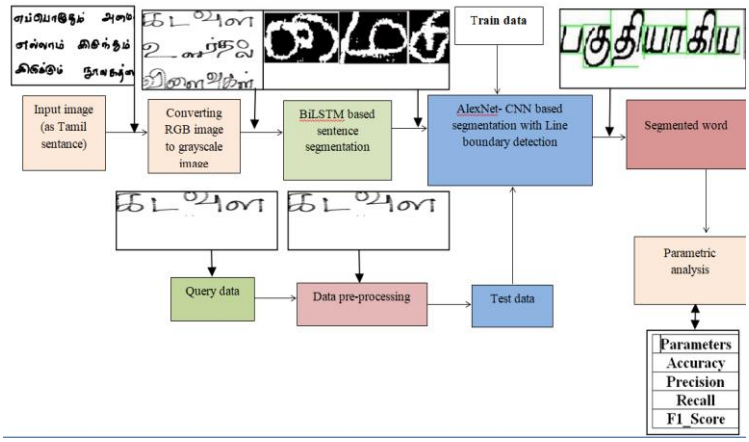


Fig.,-1 overall architecture of the proposed technique

3.1 Conversion of RGB image to Grayscale image:

In the RGB model, each tone addresses the essential shading parts Red, Green, and Blue. RGB shading pictures are addressed in the RGB shading model as red, green and blue utilizing 8-cycle

monochrome norm. Grayscale pictures are addressed by intensities. Grayscale pictures have many shades of dim in the middle of highly contrasting. The intensities of a pixel esteem is addressed inside a given reach somewhere in the range of 0 and 1(minimum and most extreme) and in the middle of fluctuating reach shades of dim which reaches is somewhere in the range of 0 and 255. 1st upsides of 3 tones (Red, Green and Blue) straight power values used in the gamma development that is prostrated below:

$$C_{\text{linear}} = \begin{cases} \frac{C_{\text{rgb}}}{12.92} & C_{\text{rgb}} <= 0.04045 \\ \frac{(C_{\text{rgb}}+0.065)}{1.065} & C_{\text{rgb}} > 0.04045 \end{cases} \quad (1)$$

Here, C_{srgb} gives RGB primary in range between zero to one and C_{linear} gives linear-intensity in range between zero to one. Luminance of outcome is gained utilizing sum of 3 linear intensity values. Conversion is gained utilizing :

$$y = f(x) \quad (2)$$

where, x gives income information and y gives outcome information. Function f(x) changes RGB measures into grayscale measures utilizing sum of R, G, and B components:

$$f(x) = 0.2989 * R + 0.5870 * G + 0.1140 * B \quad (3)$$

Illumination control, grayscale translation, Un-sharp masks, and improved adaptive thresholding for binarization are all detailed in great depth in our technique.

HSV Value is the most elevated worth among the three R, G, and B numbers. This number is partitioned by 255 to scale it somewhere in the range of 0 and 1. As far as insight, HSV Value addresses how light, splendid, or extraordinary a shading is. Esteem doesn't recognize white and unadulterated tones, all of which have V = 1. 'Esteem' is some of the time subbed with 'brilliance' and afterward it is known as HSB

The clip limit is given by: ß

$$\beta = \frac{\text{M}}{\text{N}}\left(1 + \frac{a}{100}(S_{\max} - 1)\right) \quad (4)$$

α gives clip limit factor, M gives region size, N gives grayscale value. α=100 is the Maximum clip limit.

$$I = (I_{\max} - I_{\min}) * P(f) + I_{\min} \quad (5)$$

R, G and B are 3 no., in interval [0, 255].

$$\begin{cases} H & = \cos^{-1}\left[(R - 1/2G - 1/2B)/\sqrt{R^2 + G^2 + B^2 - RG - RB - GB}\right] \text{ if } G \geq B \\ 360 & -\cos^{-1}\left[(R - 1/2G - 1/2B)/\sqrt{R^2 + G^2 + B^2 - RG - RB - GB}\right] \text{ if } B > G \end{cases}$$

We define m and M as (4) $\begin{cases} m = \min\{R, G, B\} \\ M = \max\{R, G, B\} \end{cases}$ (6)

From (4) S and V given as (5)

$$\begin{cases} \text{V} = \text{M}/255 \\ \text{S} = 1 - \text{m}/\text{M} \text{ if M} > 0 \\ \text{S} = 0 \text{ if M} = 0 \end{cases} \quad (7)$$

H, S, and V is Hue, Saturation and Value for HSV colour space.

$$\begin{cases} \text{M} = 255 \text{ V} \\ \text{m} = \text{M}(1 - \text{S}) \end{cases} \quad (8)$$

New variable is given as (7)

$$z = (M - m)\left[1 - 1\left(\frac{H}{60}\right)\%2 - 1\ |\right] \tag{9}$$

Y'UV is a colour space for coding colour pictures or movies that considers sensory perceptions. Luma (Y') is a much more suited stream for luminance estimate, depending upon its contributions to observed brightness [38][39]. It utilises combination of gamma-corrected R, G, and B. The matrix below could be used to find the connection (14)

$$\begin{bmatrix} Y' \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.144 \\ -0.14713 & -0.28886 & 0.436 \\ 0.615 & -0.51499 & -0.10001 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{10}$$

$$Y' = 0.299 * R + 0.587 * G + 0.144 * B \tag{11}$$

The Luma channel is represented by Y ′. The red, green, and blue channels are represented by R, G, and B, respectively. The Luma channel's mean average value is used to evaluate the luminance of the image. In the following phase, we'll utilise this number to calculate the increase and biases for luminance and contrasting modifications.

3.2 Otsu thresholding:

Thresholding is utilized to separate an article from its experience by relegating a power esteem T (edge) for every pixel with the end goal that every pixel is either named an item point or a foundation point. For g(x, y) is a limit adaptation of f(x, y) at some worldwide edge T, it tends to be characterized as

$$g(x, y) = \{1 \text{ if } f(x, y) \geq T \quad 0 \text{ otherwise} \tag{12}$$

Thresholding activity is characterized as:

$$T = M[x, y, p(x, y), f(x, y)] \tag{13}$$

T is the edge f(x,y) is the dark worth of point (x,y) And p(x,y) is a nearby property of the point, for example, the normal dim worth of the area fixated spot on (x, y) Converting a greyscale picture to monochrome is a conventional picture handling task. Otsu's thresholding picks the edge to limit the intraclass difference (22) of the thresholded highly contrasting pixels.

$$\sigma_w^2(t) = w_1(t)\sigma_1^2(t) + w_2(t)\sigma_2^2(t) \tag{14}$$

Observe the edge that limits the weighted inside class difference. This ends up being equivalent to expanding the between-class change

$$\sigma_b^2(t) = \sigma^2 - \sigma_w^2(t) = w_1(t)w_2(t)[\mu_1(t) - \mu_2(t)]^2 \tag{15}$$

This is communicated as far as class probabilities w_i and class implies μ_i. The class likelihood w_1 (t) is processed from the histogram as t:

$$w_1(t) = \sum_0^t p(i) \tag{16}$$

What's more the class mean μ_1 (t)is:

$$\mu_1(t) = [\sum_0^t p(i)x(i)]/w_1 \tag{17}$$

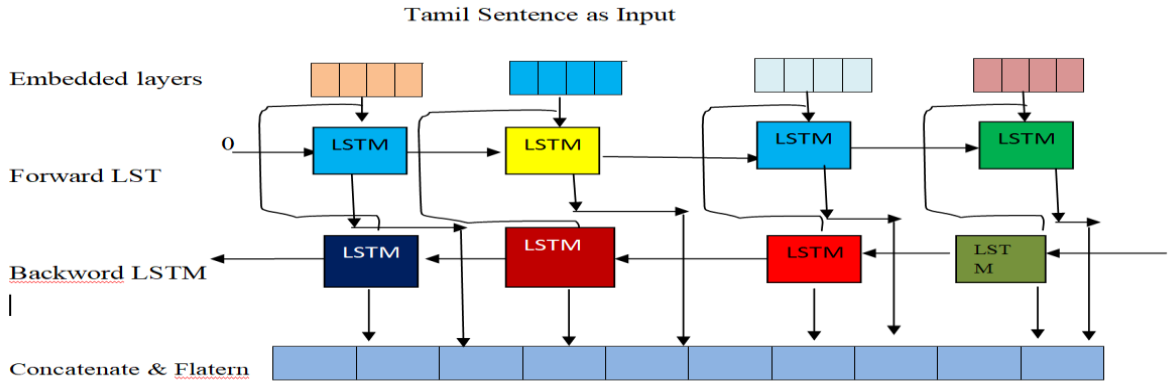
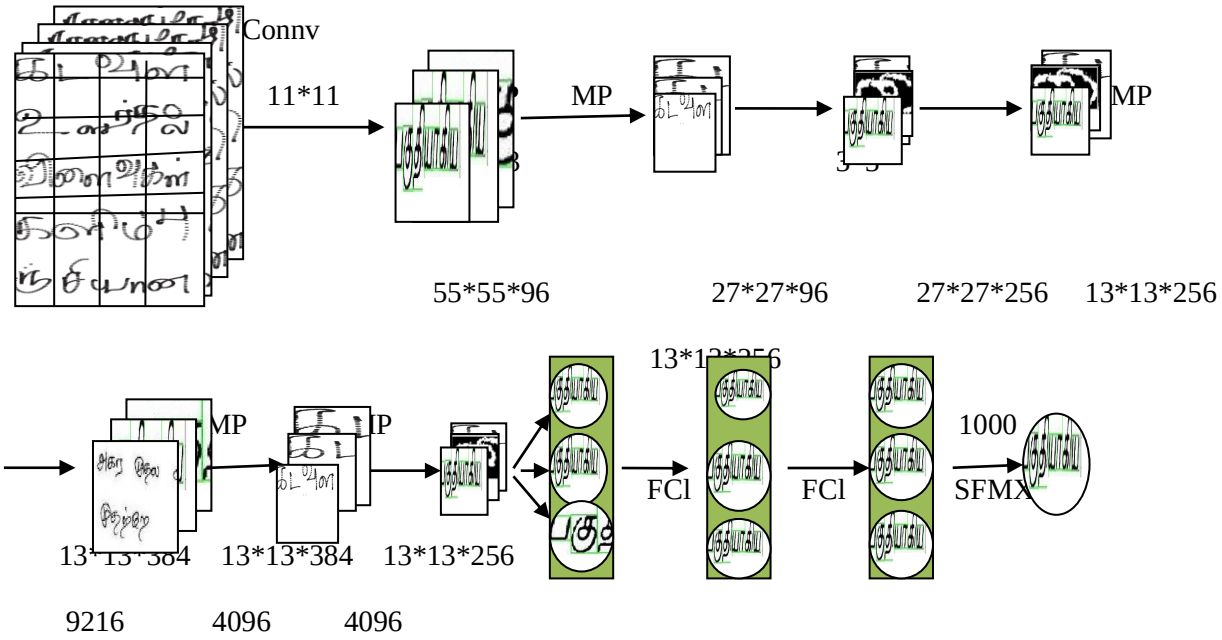


Fig., 2: Bi-LSTM architecture

3.3 Alex Net- CNN based word segmentation with Line boundary detection

To accomplish shut limits, we build a bunch of smooth bend sections, as shown by the ran bends, to associate the developed parts. Those ran bends are one more arrangement of parts. To recognize them from the underlying straight-line sections, we call them virtual pieces and call the underlying ones genuine parts. Considering the limit perfection, the virtual pieces are developed so that every one of them introduces two genuine part endpoints in G1-congruity, or at least, ceaseless areas and persistent digression bearings. Let v(t) = (x(t), y(t)), t ∈ [0, L(v)], be the bend length defined portrayal of a substantial shut limit, or at least, v(L(v)) = v(0), where L(v) is the limit length.

Convolutional Neural Networks (CNN) is a generally utilized instrument under profound learning. CNN engineering that comprises of a few layers of different sorts that are convolutional layers, actuation layers, pooling layers, and finishes with one or completely associated layers. The quantity of classes to be perceived contains similar number of neurons produce by the last completely associated layers of CNN engineering as displayed in fig.,- 4.



Fig,-3 AlexNet- CNN architecture

3.4 Performance analysis:

The whole execution of the proposed Otsu_AlexNet-CNN division is done in the Python device and the setups considered for the trial and error are: PC with Ubuntu, 4GB RAM, and Intel i3 processor.

Table- 1 Analysis of Proposed and existing techniques

Parameters	SVM	SOM	OIHACDB-28	DBN	Ostu_ alexnet-CNN
Accuracy	86	89	92	94	97
Precision	80	82	85	87	93
Recall	77	80	83	85	90
F1_Score	79	83	87	90	93

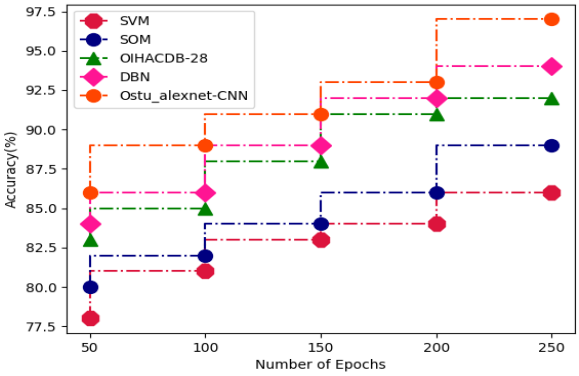


Fig.,-4 Analysis of Accuracy

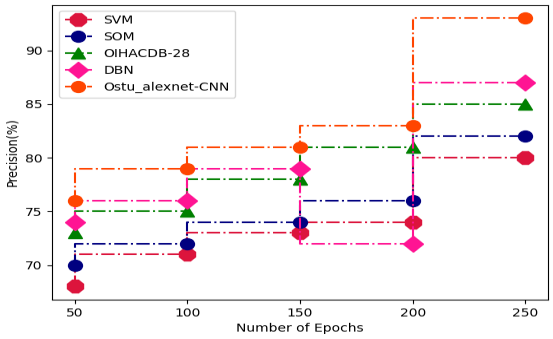


Fig.,-5 Analysis of Precision

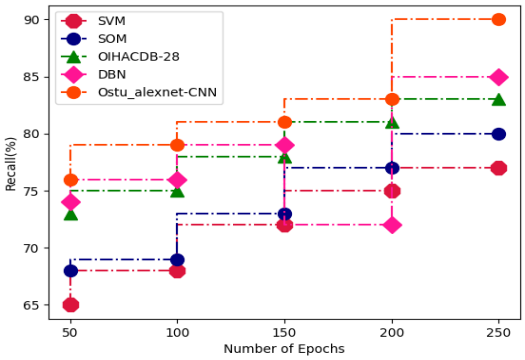


Fig.,-6 Analysis of Recall

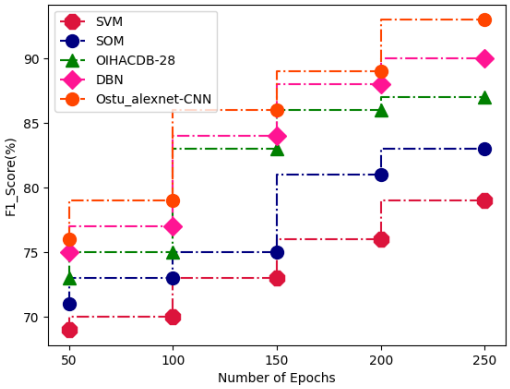


Fig.,-7 Analysis of F-1 score

The above fig., 5,6,7 and 8 shows the comparative analysis between proposed and existing techniques in terms of precision, recall, accuracy and F-1 score.

4. Conclusion:

This paper novel technique in segmenting and pre-processing the handwritten tamil word. Initially the Tamil handwritten word, the input dataset should contain Tamil handwritten word images with some image format such as BMP, JPG etc. In the image acquisition step, the documents are scanned or photograph captured. They are grey conversion, binarization, skew detection and correction, noise removal etc. Grayscale conversion and binarization is important in the handwritten word recognition to reduce the processing complexity of images. To arrange the input text and take out erroneous behaviour from that we using phases. In division step, the pre-processed files are partitioned into lines, words, and words. This paper segment the lines of handwritten word of image using Ostu thresholding with converting RGB image to grayscale image and to segment the pre-processed image using Alexnet-CNN with line boundary detection.

References:

[1] Jie Yang, Yue Zhang, and Shuailong Liang. 2019. Subword encoding in lattice LSTM for Chinese word segmentation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long and Short Papers). 2720–2725.

[2] Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Transition-based neural word segmentation. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 421–431.

[3] Qi Zhang, Xiaoyu Liu, and Jinlan Fu. 2018. Neural networks incorporating dictionaries for Chinese word segmentation. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence.

[4] Hai Zhao, Deng Cai, Changning Huang, and Chunyu Kit. 2019. Chinese word segmentation: Another decade review (2007- 2017). arXiv preprint arXiv:1901.06079 (2019).

[5] Lujun Zhao, Qi Zhang, Peng Wang, and Xiaoyu Liu. 2018. Neural networks incorporating unlabeled and partially-labeled data for cross-domain chinese word segmentation. In Proceedings of the International Conference on Artificial Intelligence (IJCAI'18). 4602–4608.

[6] Hao Zhou, Zhenting Yu, Yue Zhang, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2017. Word-context character embeddings for Chinese word segmentation. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 760–766

[7] Bannigidad, P. and Gudada, C. Age-type identification and recognition of historical kannada handwritten document images using HOG feature descriptors. In Computing, Communication and Signal Processing, (pp. 1001- 1010). Springer, Singapore, (2019).

[8] Neche, C., Belaid, A., & Kacem-Echi, A. (2019, September). Arabic handwritten documents segmentation into text-lines and words using deep learning. In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW) (Vol. 6, pp. 19-24). IEEE.

[9] He, H., Wu, L., Yan, H., Gao, Z., Feng, Y., & Townsend, G. (2019). Effective neural solution for multi-criteria word segmentation. In Smart Intelligent Computing and Applications (pp. 133-142). Springer, Singapore.

[10] Hellwig, O., & Nehrdich, S. (2018). Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In Proceedings of the 2018 conference on empirical methods in natural language processing (pp. 2754-2763).