**Signboard Translation From Vernacular Languages**

### A PROJECT REPORT

***Submitted by***

**VISHNU TEJA CHIKKALA [RA1711003030378]**

**RAJPREET SRIVASTAV [RA1711003030380]**

*Under the guidance of*

## MR. Sunil Kumar

(Professor, Department of Computer Sciene & Engineering)

***in partial fulfillment for the award of the degree of***

### BACHELOR OF TECHNOLOGY

in

## COMPUTER SCIENCE & ENGINEERING

of

### FACULTY OF ENGINEERING AND TECHNOLOGY



#### SRM INSTITUTE OF SCIENCE & TECHNOLOGY, NCR CAMPUS

**MAY 2021**

SRM INSTITUTE OF SCIENCE & TECHNOLOGY

(Under Section 3 of UGC Act, 1956)

## BONAFIDE CERTIFICATE

#### Certified that this project report titled **Signboard Translation from Vernacular Languages** is the bonafide work of **VISHNU TEJA CHIKKALA [RA1711003030378], RAJPREET SRIVASTAV [RA1711003030383]** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form any other project re- port or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

##### SIGNATURE

MR.SUNIL KUMAR

##### GUIDE

Professor

Dept. of Computer Science & Engineering

Signature of the Internal Examiner

##### SIGNATURE

Dr. R.P.MAHAPATRA

##### HEAD OF THE DEPARTMENT

Dept. of Computer Science & Engineering

Signature of the External Examiner

# ABSTRACT

India has 22 constitutionally recognized languages written in 13 different scripts. An average traveler, on a business trip, when travelling to a new region, might often get confused with signboards written in an unfamiliar language. It is also impossible to have every signboard in every city / town / village written in 22 different languages, as there will not be enough room to accommodate more than 2 – 3 scripts.

The objective of this project is to develop a simple, easy-to-use and scalable Android mobile app which provides a two-click, picture-to-text, translation / transliteration service for Indian vernacular languages, using deep neural networks and natural language processing models trained for text detection, recognition and translation tasks on collected and freely-available datasets, using frameworks such as PyTorch and Tensorflow.

For the scope of this project, we will design a system which works for names (such as road names, city names, shop names, organization names, etc.) which typically are not longer than 4-5 words, and support translation for 1 language. Further scope for the project involves including support for more languages and building models to support translation / transliteration of longer pieces of text

**ACKNOWLEDGEMENT**

We would like to express our deepest gratitude to our guide, **MR. SUNIL KUMAR** for his valuable guidance, consistent encouragement, personal caring, timely help and providing us with an excellent atmosphere for doing research. All through the work, in spite of his busy schedule, he has extended cheerful and cordial support to us for completing this research work.

##### Authors

**TABLE OF CONTENTS**

**LIST OF TABLES**

**LIST OF FIGURES**

**ABBREVIATIONS**

**ANN** - Artificial Neural Network

**CNN** - Convulutional Neural Network

**RNN -** Recurrent Neural Network

**GRU -** Gated Recurrent Unit

**CTC** - Connectionist Temporal Classification

**NITS-LD** - NIT Silchar Language Database

**OGI-MLTS** - OGI-Multilingual Database

**CHAPTER I**

**INTRODUCTION**

* 1. **Project Overview**

The project describes the process of creating an end-to-end stack of neural network models, to perform the tasks of text detection, text recognition and text transliteration in order to extract text of the source language from an image, and output transliterated text in the target language. For this purpose, three models are built:

* a Region-based Convolutional Neural Network for performing text detection (a subset task of object detection with class 'text')
* a Convolutional Neural Network (CNN) - Recurrent Neural Network (RNN) / Gated Recurrent Unit (GRU) Encoder-Decoder Architecture for performing text recognition
* a GRU - GRU Encoder-Decoder Architecture for performing text transliteration

This models will then be trained on a custom dataset featuring a large number of synthetically generated scene-text images, as well a collection of natural scene-text images, along with annotations describing bounding boxes and ground truth source script. The transliteration model will be trained on another dataset which includes large number of source script words and their corresponding target script words.

This model will then be compiled into an end-to-end stack using TensorflowLite and deployed onto an Android app developed in Java.

**CHAPTER II**

**LITERATURE REVIEW**

**2.1 Summary of Literature Review**

C.Kurian et al. concluded that the Indian community faces a “Digital Divide” due to dominance of English as mode of communication in higher education, judiciary, corporate sector and Public administration at Central level whereas the government in states work in their respective regional languages[1].

From the 2011 Census of India, it is observed that. while 99 % of the population speak one of these scheduled languages in various dialects (which number in the thousands), according to Census 2011, the total percentage of English speakers is at 10 %, and that too is skewed towards the urban population Hence, there lies a need for developing NLP architectures for facilitating flflow of digital content and information in and between local, national and international levels[2].

N.P. Desai et al noted that traditionally NLP had been approached with statistical methods such as Hidden Markov Machines (HMM), Support vector machine(SVM), Conditional Random Field(CRF), Naive Bayes(NB), etc, which take a large amount of tagged/annotated data (corpus) to statistically analyze and learn the language characteristics , and suggested that deep learning methods or a ‘connectionist approach’ could return better results[3]. The reasons for the same suggested by them and other authors, in brief, were -

1. the simplicity of the solution in rapidly prototyping and establishing practically effffective systems
2. the lower cost of annotation of the training data [4],
3. they attempt to more closely emulate the learning process of biological brains [3] [5] [6], among other reasons.

However, A. Kunchukuttan et al., note that the collection of a uniform corpus and standard datasets for training models remains a challenge across all regional languages. The large number of morphological variations across Indic languages also contributes to this issue[7]. This study also proposed the creation of a “large-scale, general-domain” corpus for 10 Indian Languages across 2 scripts[7]. J. Philip et al., in their study, also highlighted several efforts at collating standard datasets and corpora for Indian languages[4].

Sharma et al., 2017 concluded that almost all existing Indian language machine transliteration systems are based on statistical and hybrid approach[8].

A few models proposed in this field are as follows-

|  |  |  |
| --- | --- | --- |
| **Authors** | **Model** | **Accuracy** |
| Bhanja et al., 2019 | CNN-LSTM Architecture using ResNet | 93% (NITS-LD), 89% (OGI-MLTS) |
|  |  |  |

**REFERENCES**

1. C. Kurian and K. Kannan Balakrishnan.(2008). “Natural language processing in india prospects and challanges”.*In Proceedings of the International Conference on “Recent Trends in Computational Science.*
2. ORGI. ["Census of India: Comparative speaker's strength of Scheduled Languages-1951, 1961, 1971, 1981, 1991 ,2001 and 2011"](http://www.censusindia.gov.in/2011Census/C-16_25062018_NEW.pdf)
3. N. P. Desai and V. K. Dabhi. (2021) “Taxonomic survey of hindi language NLP systems”, *arXiv preprint arXiv:2102.00214,*
4. J. Philip, V. P. Namboodiri, and C. Jawahar (2019), “A baseline neural machine translation system for indian languages”, *arXiv preprint arXiv:1907.12437*.
5. M. Rosca and T. Breuel. (2016) ”Sequence-to-sequence neural network models for transliteration.”. *arXiv preprint arXiv:1610.09565*.
6. T. Deselaers, S. Hasan, O. Bender, and H. Ney (2009). “A deep learning approach to machine transliteration”. *In Proceedings of the Fourth Workshop on Statistical Machine Translation, pages 233–241.*
7. A. Kunchukuttan, D. Kakwani, S. Golla, A. Bhattacharyya, M. M. Khapra, P. Kumar, et al. (2020). “Ai4bharat-indicnlp corpus: Monolingual corpora and word embeddings for indic languages.” *arXiv preprint arXiv:2005.00085.*
8. A. Sharma and D. Rattan.(2017) *“*Machine transliteration for indian languages: A review*.” International Journal of Advanced Research in Computer Science, 8(8),*