International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 04 | Apr 2020 www.irjet.net p-ISSN: 2395-0072

Language Linguist using Image Processing on Intelligent Transport Systems

Sampada Gaonkar[1], Anubhuti Rane[2], Gauri Gulwane[3], Tamanna Kasliwal[4], Dr. Chaya Jadhav[5]

[1,2,3,4]Student, Dept. of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Maharashtra, Pune
[5]Associate Professor, Dept. of Computer Engineering, Dr. D. Y. Patil Institute of Technology, Maharashtra, Pune

---

Abstract - *Visitors traveling to various countries around the world often find it difficult to understand and communicate the local languages, since they do not know it. Within the new places they cannot read the words written on the boards or banners. Therefore, text extraction systems need to be built which can identify and recognize text found in the navigation board .The system proposes a three stage process that involves detection, extraction and translation using the concepts of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). The framework is designed to take into account the need to create a desktop application that extracts the text from images based on traffic navigation boards and translates it further into a user-friendly language. By this way the user can grasp the unfamiliar language quickly.*

Key Words: Detect, Extract, Translate, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN).

1. INTRODUCTION

Travelling to new places always involves difficulty in navigating to the desired destination when the language used for communication is not understandable to the person. Thus, this poses a problem and limits the travelling experience. Also, the various applications available wherein the user has to type the text in order for it to be translated to that user's language is a tiresome job. Hence, proposing a system that can easily detect and extract the text from images and further translate where the user only needs to upload the image.

In this system, it aims to build a desktop application that helps translating text in images from Spanish to English or from French to English. In artificial intelligence (AI) and natural language processing (NLP), machine translation is an important research topic. Along with growing tourism, people find it difficult to understand the native language of the place they are travelling to. Since, they are unable to interpret the words written on the navigational boards or banners, the system is developed that will extract the text from images uploaded by the user. Furthermore, the extracted text would be translated to English language, and help the users to understand the navigational boards.

In addition, text-based traffic signs in countries like Spain and French usually contain Spanish language and French language respectively. However, so far there is no unified method to deal with the text-based traffic signs in various languages. Therefore, text-based traffic sign detection and further translating it to English language is still a very challenging task. In recent years, with the continuous success of deep neural networks in many fields, they have become the mainstreams for many vision tasks. The proposed system aims to investigate how to use the deep learning tool to solve the problem of text-based traffic sign detection, extraction and translation. The goal is to accurately detect the texts in traffic signs with high efficiency, fully avoiding the influence of background texts and symbol-based traffic signs.

1.1 OBJECTIVE:

The system resolves to solve the issues faced by the travellers that have trouble understanding the local language of the place they are visiting. In this case, the language is Spanish or French.

1) To detect the text area from the image consisting of text-based navigation boards by eliminating the non-textual areas.
2) To extract the words from the identified text area.
3) To translate the extracted text into user understandable language, that is, English

2. LITERATURE SURVEY

Yingying Zhu et al. [1]. The traffic based signs mainly consists of traffic signs or text based navigational boards. To locate the signs from the images, the detection process is a two-stage detection method that reduces the search area of text detection and removes texts outside traffic signs. The language of the text based traffic signs are in English and Chinese which are trained based on a public dataset and the self- collected dataset. This system makes use of fully convolutional neural network to train the images. The proposed application improves the detection speed and solves the problem of multi-scale for text detection. Traffic sign detection and text detection are two different object detection problems, it is not reasonable to detect two different objects in a unified framework. Also, it is hard for text-boxes to detect the texts because they are too small in the whole images.

Youbao Tang et al. [2] presents text detection and segmentation using cascaded convolution network (CNN). Candidate text region (CTR) extraction model is created

using edges and the whole region. The CTR is then converted to text then to refined CTRs. The refined CTR is classified using CNN based CTR classification model (Cnet). This system generates more precise text region which do not have much background noises than the traditional techniques. This system needs the decisive information of text edges and regions to train the models. But the publicly available datasets which contain this decisive information are too small to effectively train these models hence it is hard to train these datasets.

Yongchao Xu et al. [3] presents a text detector called TextField for detecting irregular scene texts. A direction field is learned and using this direction field text is detected fully using cascaded convolution network (CNN). A morphological-based post-processing is applied to learning direction to get the final detection. This system provides grouping of the text regions and gives better performance and efficiency. The traditional segmentation is challenging than this method as it hardly separate adjacent text instances. It still fails for some difficult pictures, including occlusion of objects, spacing of large characters. TextField also has some incorrect detections on certain text-like regions.

Zhaorong Zong et al. [4] presented that the source language is distorted as it passes through a noisy channel and appears as Target language at the other end of the channel. The task of the statistical machine translation system is to find the highest probability of the sentence as a translation result. Neural machine translation has replaced statistical machine translation. The idea of end-to-end neural machine translation is to directly implement the automatic translation between natural languages through neural networks. For this, neural machine translation usually uses an encoder-decoder framework and achieves sequence-to-sequence conversion. After that, the target language uses another recursive neural network to reversely decode the source language sentence vector to generate the target language. This entire process of decoding is generated word by word. Hence, the decoder can be regarded as a language model containing the target language of the source language information. Therefore, the encoder-decoder model based on attention mechanism changes the way of information transmission and can dynamically calculate the most relevant context so as to better solve long-distance information transmission problems and improve the performance of neural machine translation.

Jack Greenhalgh et al. [5] presented a method of detection and recognition of traffic signs which runs at a frame rate of 14 frames/s, under Open Source Computer Vision (OpenCV), on a 3.33-GHz Intel Core i5 CPU. By this method, considerable increase in speed was gained by running the algorithm in parallel as a pipeline. Their proposed system consists of two main stages: detection and recognition. Detection consists of three phases: determination of search regions, identifying large numbers of text-based traffic sign candidates using basic color information and shape and reduction of candidates using contextual constraints. This over detection is important to make sure that no true positives (TPs) are missed. Potential candidate regions for traffic signs are then located only within the scene search regions, using a combination of MSERs and hue, saturation, and value (HSV) color thresholding. Then these large number of detected candidate regions are reduced by making use of the structure of the scene and temporal information, to eliminate unlikely candidates. Next step is recognition. Candidate components for text characters are then located within the region and sorted into potential text lines, before being interpreted using an off-the-shelf optical character recognition (OCR) package. The set of detected text lines (in grayscale) are passed on to the open-source OCR engine "Tesseract" for recognition. To improve the accuracy of OCR, results are combined across several frames to improve the accuracy of recognition.

Fares Aqlan et al. [6] proposes the methodology to translate the complex language like Arabic to Chinese and vice-versa, which serves as a difficult task due to large number of rare words. The techniques byte pair encoding (BPE) and Neural Machine Translation (NMT) are used where the rare words can be effectively encoded as sequences of sub-word that includes Romanization. The method also provides a qualitative analysis of the translation results. It also compares the impact of various segmentation strategies on Arabic-Chinese and Chinese-Arabic NMT system while proposing the standard criteria for data filtering of Chinese-Arabic parallel corpus. The results of the translation can be more accurate and efficient by adopting other languages to increase the segmentation consistency between Romanized Arabic and those languages.

Kehai Chen, Rui Wang et al. [8] presents the sentence-level context as latent topic representations by using a convolution neural network(CNN), and designs a topic attention to integrate source sentence-level topic context information into both attention-based and Transformer-based NMT. This method can improve the performance of NMT by modelling source topics and translations jointly. It is a variant of CNN that captures source topic information based on the sentence-level context. The proposed CNN maps source topic information implicitly into topic vectors and called as latent topic representations (LTRs).

Tiejun Zhao et al. [7] presents source dependence-based context representation for translation prediction. This neural network approach is to encode bilingual context. It is capable of not only encoding source long-distance dependencies but also capturing functional similarities to better predict translations. This method can significantly improve SMT performance over strong base- line methods, and verified that structural clues in context were beneficial for translation prediction.

Muhammad A.Panhwar et al.[9] presents that machines learning and pattern recognition play a vital role in extracting information from natural scenes. The system

introduces a framework to extract the text from natural landscape and natural scenes. Capturing image is the very first step for signboard detection. The next step is real-time signboard detection followed by Text detection and Recognition. In recognition, the image text (pixel-based text) is converted to a readable and editable form. The system performed experiments on 500 images randomly from natural scenes. Neural Network has been selected for the recognition of these images. The system performs recognition for both English and Urdu languages. The system achieved up to 85% accuracy in signboard detection.

3. BASIC CONCEPTS

Machine Translation (MT) is a sub-field of computational linguistics that focuses on translating text from one language to another. Neural Machine Translation (NMT) has emerged as the most effective algorithm, with the aid of deep learning, to perform this task.

a) Convolutional Neural Network

Convolutional Neural Network (CNN) supporting the principle of deep learning is the key focus in Artificial Intelligence, which can take on an input image, significance (learnable weights and biases) for different aspects / objects in the picture to recognise from one to another. Object classification is the process of taking input of an object and outputting a class or class probability that best describes the image. The CNN techniques help to identify patterns easily, generalize from prior experience and adjust to various picture environments. A more detailed explanation of what CNNs are doing is taking the picture, moving it through a series of nonlinear, convolutional, pooling and completely connected layers, and generating an output.

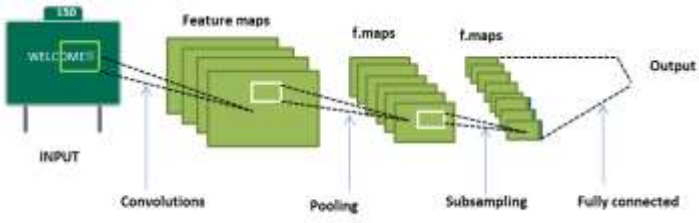

Fig 1: Structure of Convolutional Neural Network (CNN)

The first layer is often a convolutional layer. A filter neuron in CNN is a collection of learnable weights that the back-propagation algorithm is used to train. It can store a template / pattern. A very significant point is that this filter depth needs to be equal to the depth of the input. In a filter, the values are multiplied as it converges around the image to give the final single number as multiplications are all added. Every single position on the input volume generates a number. Hence, each of these filters can be addressed as feature identifiers. Another building block to a CNN is a pooling layer. The purpose is to slowly reduce the representation's spatial size in order to minimize the amount of parameters and computation within the network. Pooling layer operates on each feature map independently. Max pooling is the most common method employed in pooling. A Fully Convolutional Network (FCN) is one wherein all the learning layers are convolutional, and there is no fully linked layer in it. A fully convolutional neural network resembles a convolutional neural network with no entirely connected layers ie- It consists of strictly convolutional layers, and probably some max-pooling layers. This means that the network output is an image as the output layer is a convolutional layer. Thus, the model is trained based on these convolutional layers which generate the desired output.

b) Recurrent Neural Network

Recurrent Neural Network (RNN) falls under a special category of neural network with loops that allow information to persist in a network over various steps. It can also be thought of as constantly utilizing the same network, with each new addition the model has a little more knowledge than the previous one.

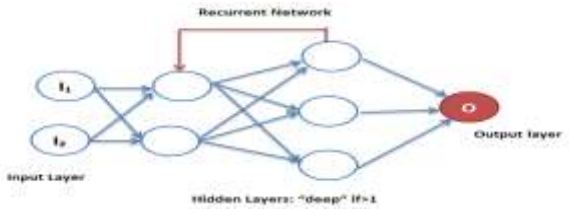

Fig 2: Working of Neural Network (RNN)

Although RNNs learn similarly during training, it often remember things learned while generating output from prior data. It is a part of the network. RNNs can take one or more input vectors and generate one or more output vectors, and the output(s) are determined not only by weights applied to inputs such as a normal neural network, but also by a "hidden" state vector representing the background based on prior input / output

So, depending on previous inputs in the sequence, the same input may produce a different output. The inputs are processed sequentially by a recurrent neural network. A recursive neural network is similar to the degree the transformations are applied to inputs repeatedly, but not necessarily in a sequential fashion. Recursive Neural Networks are a more general type of Neural Networks. It can function on any structure of the hierarchical tree. Unlike feedforward neural networks, input sequences can be processed by RNNs using their internal state (memory). Over time, an RNN remembers any and every detail. In time series prediction it is only useful to remember previous inputs even because of the function. Therefore, this is known as Long Term Short Memory.

4. SYSTEM ARCHITECTURE

I. Text Detection and Extraction Module

The aim is to perform text detection on a text-based navigational board based image for detection and extraction of text. In the first step, the detection of rectangular regions is done that potentially contain text. In the second step, the system performs text extraction, where, for each of the detected regions, a Convolutional Neural Network (CNN) (detailed explanation in section 3. BASIC CONCEPTS) is used to recognize and decipher the word in the region.

This module can further divided into:

a) Detection of text areas from non-textual areas in images.

b) Extraction of the text area.

c) Presentation of the detected text in standard format.

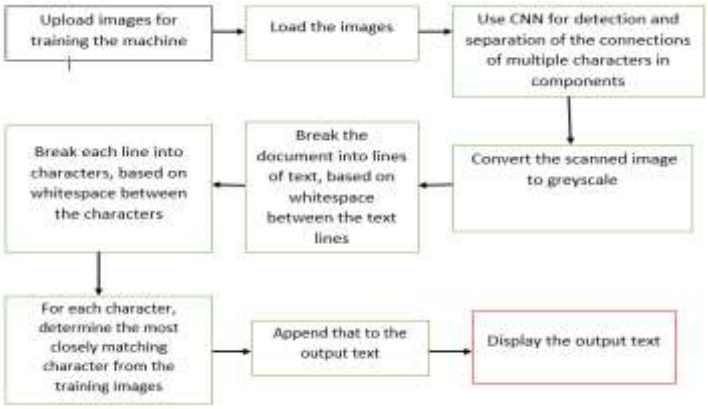

Fig 3: Text Detection and Extraction Architecture

a) Detection of Text Areas from Non-Textual Areas in Images:

The images obtained in this domain comprise of traffic navigational boards. Thus main motive is to obtain the text from these images. The entire detection system is jointly trained in a controlled, end-to-end manner. The users can upload the images that they want to be translated through the user interface provided by the application. The use of a learning rate scheduling is done, starting with a very low learning rate to ensure that the model doesn't diverge, and progressively increase the learning rate during the first few epochs to ensure a nice, stable point is reached in the model.

Grayscale images are composed of only grey tones of colour, which are of 256 steps. There are, in other words, only 256 grey colours. The key feature in grayscale pictures is the amount in the levels of red, green, and blue. The color code would be as RGB(R, R, R), RGB(G, G, G), or RGB(B, B, B) where 'R, G, B' is a single digit between 0 and 255.

Ultimately, the image is composed of a number of pixels, and these pixel values can differ from one range to another.

1) The pixel value for binary image is either 1 or 0 which means only two shades 1=white and 0=black.

2) In the case of gray image scale, for example 8-bit gray image scale (2 ^ 8=256) pixel value can vary from 0 to 256. Here the image is 256 shades (0=black, 1=white, and for others the combination of both).

Hence, Grey Scale image can have shades of grey differing between Black and white while Binary image can either of two extreme for a pixel value either white or black. This makes it easy for the grayscale images for the removal of noise in images while preserving edges.

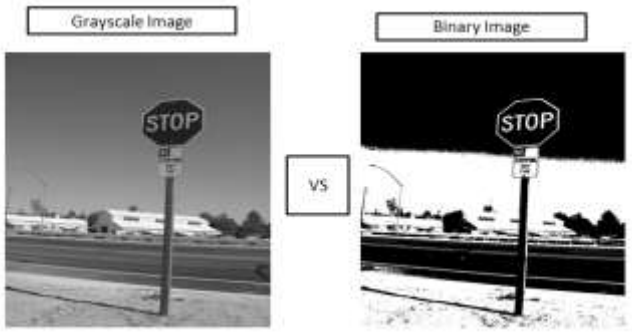

Fig 4: Grayscaling vs Binary Image

This detection process is carried out using Convolutional Neural Network, where each algorithm that can take in an input image, assign importance to different aspects / objects in the image, and be able to distinguish between them. For CNN, the pre-processing level is much lower compared with other forms of classification. This algorithm is best used in the identification of artifacts to distinguish patterns, edges and other variables such as colour, strength, shapes and textures.

b) Extraction of the Text Area:

Textboxes are areas of importance where detections from all the irrelevant sections of the picture are the text field. The textboxes can also serve as regions that are essential for this type of model, since determining the location of multiple objects. After detections of regions in the image, the following steps are done:-

1) Dividing each line into characters, built on whitespace between the characters.

2) Divide the document into lines of text, built on whitespace between the lines.

c) Presentation of the Detected Text in Standard Format:

1) Determine the most closely matching character from the training model for each character extracted from the Text field.

2) Append the identified character in the output text.

3) Display the generated text in the standard format.

II. Text Translation Module:

The Stage I process in translation depicts the pre-processing steps where the datasets of Spanish, French and English is imported through the setup of libraries. The pre-processing stage comprises to clean text by removal of noise to obtain accuracy. The steps such as removal of non-printable characters and symbol based characters like punctuation marks. To normalise the Unicode characters to ASCII (American Standard Code for Information Interchange) and finally the elimination of tokens that are not alphabetic. The final output is a noise free text used for translation process

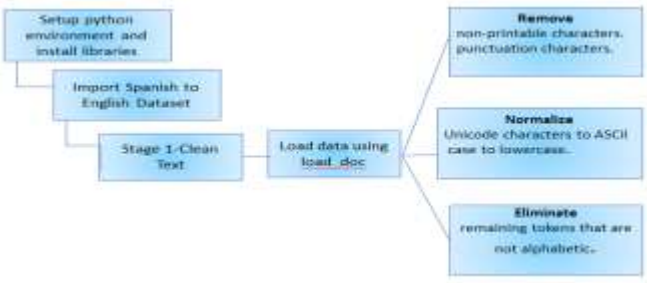

Fig 5: Stage I of Text Translation Process

The Stage II process of the translation is based on the training and testing of the model. The datasets are spilt upon the necessary criteria of splitting into train and test sets. The model uses the concept of tokenization wherein the words identified are attached with a unique numerical value which remains constant throughout. The model is mainly dependent upon the Recurrent Neural network (RNN) (detailed explanation in section 3. BASIC CONCEPTS) and Long Short-Term Memory process (LSTM).Thus, the model is trained on these basis. The trained model acts upon the test set or given input from the system where it finally translates the text from Spanish to English or from French to English.

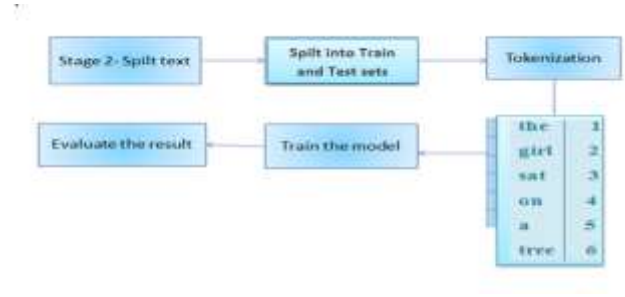

Fig 6: Stage II of Text Translation Process

5. CONCLUSION

The system proposes a technique with Convolutional Neural Network and Recurrent Neural Network to produce the best translated text after it is extracted from the detected navigational boards. This system therefore detects and extracts texts from the Spanish and French navigational boards and translates them into the user understandable language which is English. It will thus overcome the difficulties faced by the travellers who are unable to understand the unknown languages.

REFERENCES

[1] Yingying Zhu, Minghui Liao, Mingkun Yang, and Wenyu Liu. "Cascaded- Segmentation- Detection Networks for Text-Based Traffic Sign Detection" IEEE Transactions On Intelligent Transportation Systems, 2018.

[2] Youbao Tang and Xiangqian Wu, "Scene Text Detection and Segmentation Based on Cascaded Convolution Neural Networks" IEEE Transaction on Image Processing, 2017.

[3] Yongchao Xu, Yukang Wang, Wei Zhou, Yongpan Wang, Zhibo Yang, Xiang Bai. "TextField: Learning A Deep Direction Field for Irregular Scene Text Detection" IEEE Transaction on Image Processing, 2019.

[4] Zhaorong Zong, Changchun Hong. "On Application of Natural Procesing in Machine Translation" 3rd International Conference on Mechanical Controland Computer Engineering, 2018.

[5] Jack Greenhalgh, Majid Mirmehdi. "Recognizing Text-Based Traffic Signs" IEEE Transaction on Intelligent Transportation Systems, 2014.

[6] Fares Aqlan, Xiaoping Fan, Abdullah Alqwbani, and Akram Al-Mansoub. "Arabic-Chinese Neural Machine Translation: Romanized Arabic As Subword Unit For Arabic-Sourced Translation" IEEE Access, 2019.

[7] Kehai Chen , Tiejun Zhao, Muyun Yang, Lemao Liu , Akihiro Tamura , Rui Wang , Masao Utiyama,and EiichiroSumita. "A Neural Approach to Source Dependence Based Context Model for Statistical Machine Translation" IEEE/ACM Transactions On Audio, Speech, And Language Processing, 2018.

[8] Kehai Chen, Rui Wang, Masao Utiyama, EiichiroSumita, and Tiejun Zhao. "Neural Machine Translation with Sentence-level Topic Context" IEEE/ACM Transactions on Audio, Speech, And Language Processing, 2019.

[9] Muhammad A.Panhwar Kamran A. Memon, Saleemullah Memon Sijjad A. Khuhro. "Signboard Detection and Text Recognition Using Artificial Neural Networks Signboard Detection and Text Recognition Using Artificial Neural Networks" IEEE Transcation on Image Processing, 2019.