Text Recognition

Than Lwin Aung

Text Recognition

Text Recognition is one of the most difficult and challenging tasks in Artificial Intelligence and Machine Learning.

Of course, recognition of a single character is not that difficult; however, to recognize a text in all possible situations, known as **Scene Text Recognition** (STR), is not at all a trivial task.

Ambient lights, lens distortions, background noises, textures or different text styles and orientations, can have huge impacts on the recognition of a text.

Although it is commonly known as Text Recognition, it actually has 2 major steps in any Modular Text Recognition System: **Text Detection** and **Text Recognition**.

Indeed, there are end-to-end Text Recognition Systems. However, our method employs a modular approach.

Scene Text Recognition

Source: www.researchgate.net

Text Detection

Even before actually recognizing a text, it is even more important to locate (spot) where the text is.

In fact, the principal ideas of Text Detection come from Object Detection.

The earlier implementations of Text Detection (Spotting) is based on Object Detection Methods, such as SSD (Single Shot Detector).

One of the most famous model of Text Detection, which is based on SSD, is TextBox++.

Although Text Detection can be assumed to be similar to Object Detection, there is a fundamental difference: A Text is only detected to be read (recognized).

Text Detection

There are 2 primary models of Text Detection: Region Level Detection and Pixel Level Detection.

Region Level Detection primarily comes from the ideas of Object Detection like SSD, which usually employs CNN (Convolutional Neural Network).

Also, Pixel Level Detection comes from the ideas of (Semantic) Image Segmentation, which usually employs FCN (Fully Convolutional Network).

There are also many famous models for Text Detection based on Pixel, such as EAST.

Text Detection

In fact, Text Detection is only the first stage of Text Recognition, where spotting a text is not enough as it is essential to be able to be read.

Therefore, Pixel Level Detection has more advantages as Legible Text is very helpful in Text Recognition.

Most Pixel Level Text Detection employs Binary Classification of Pixels, along with additional information such as Bounding Boxes.

Binary Classification of Pixel is represented by “Text” or “No Text” Classification.

However, another Pixel Level Detection Model, known as CRAFT (Naver Inc.), employs Gaussian Heatmaps to classify Text Region (Pixel).

Character Region Awareness of Text Detection

Reference: From Original CRAFT Paper.

Figure 1. Visualization of character-level detection using CRAFT. (a) Heatmaps predicted by our proposed framework. (b) Detection results for texts of various shape.

Character Region Awareness of Text Detection

Reference: From Original CRAFT Paper.

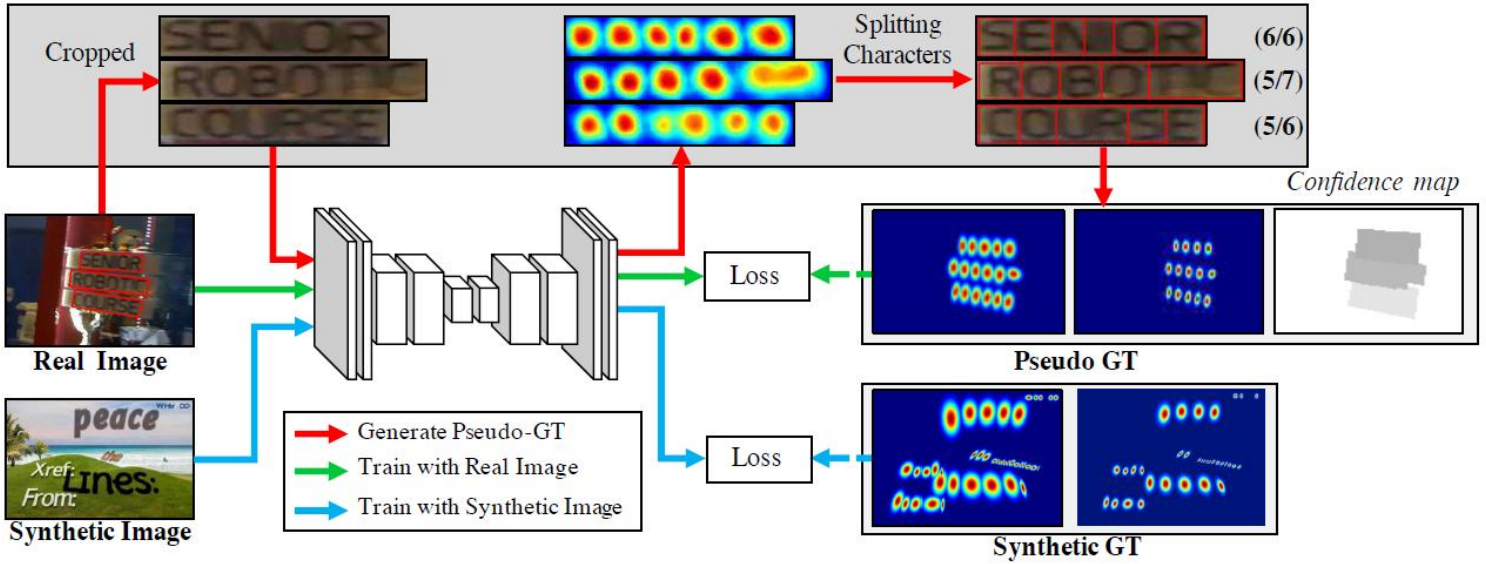


Figure 4. Illustration of the overall training stream for the proposed method. Training is carried out using both real and synthetic images in a weakly-supervised fashion.

Scene Text Dataset Generation

Our Text Detection Model fully employs CRAFT Model.

However, Detecting Burmese Characters is not a trivial task.

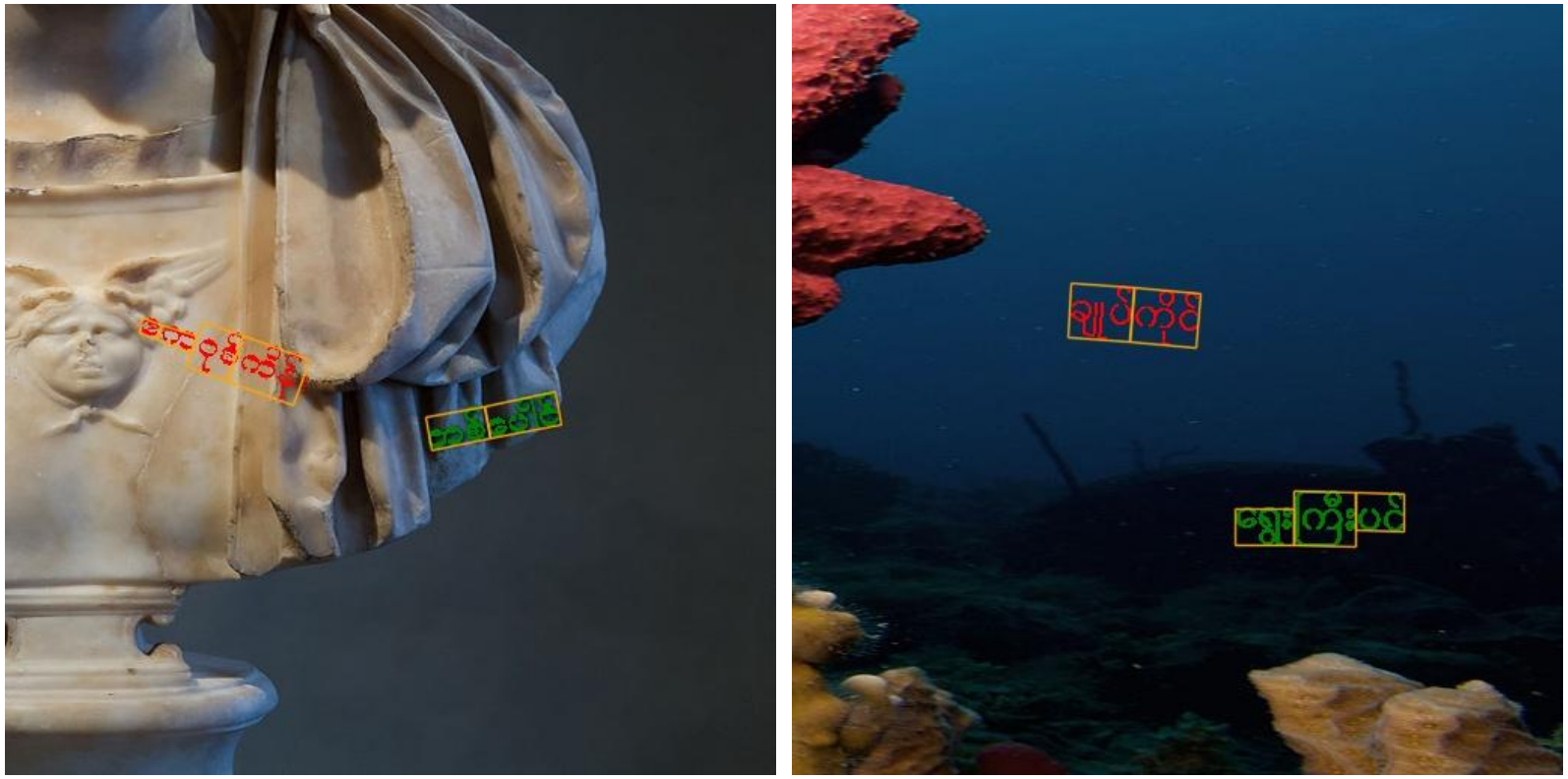
One of the major problem we have encountered is **Burmese Scene Text Dataset**.

Since there is almost none Burmese Scene Text Dataset, we have to create our own Scene Text Dataset, with Scene Text Generation.

In addition, CRAFT Model requires Texts to be segmented by Character Level.

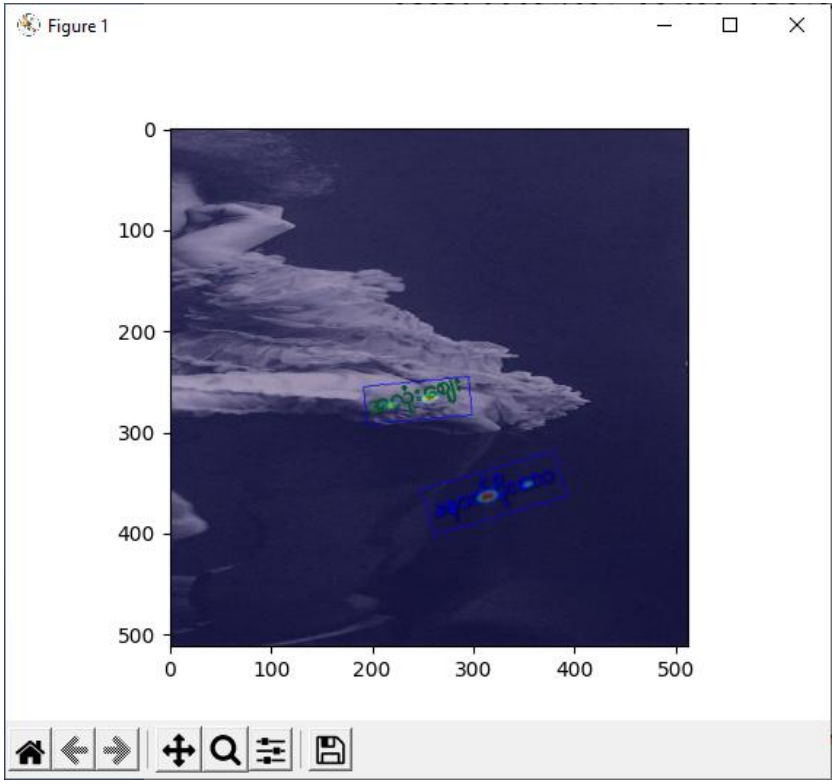
Therefore, we come up with a novel idea: **Syllable Level Burmese Scene Text Generation**.

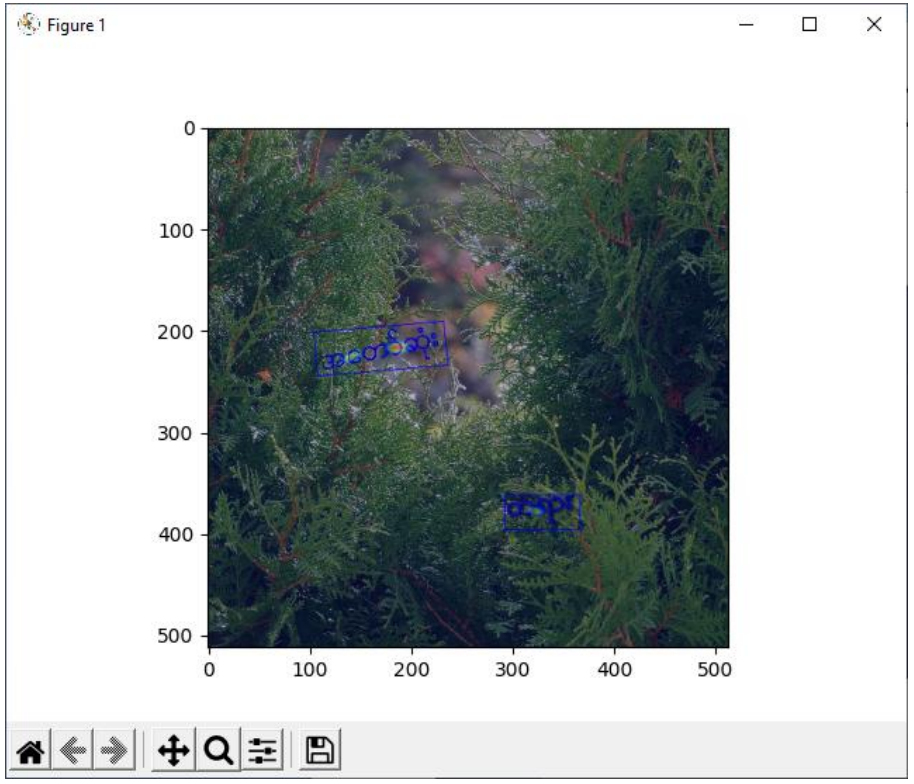
Syllable Level Burmese Scene Text Generation



Generated Burmese Text Images are segmented into Syllable Level Bounding Boxes.

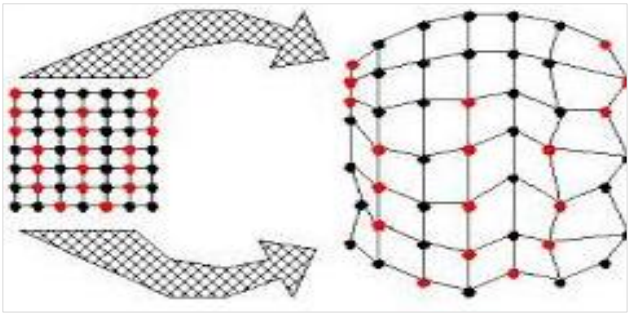
Burmese Text Detection with CRAFT





With CRAFT, it is also possible to detect Word Level Burmese Text, with Bounding Box.

Warp Transformation



Texts in Real World are rarely in Recognizable Form; it is more likely that they will be rotated in arbitrary direction or will be deformed with Perspective.

In order to solve these problems, Warp Transformation is necessary to extract the detected Text from Text Detection Model.

In addition, we can employ STN (Spatial Transformer Network) to further alleviate Text Deformation.

However, in our case, simple Warp Transformation seems to be enough.

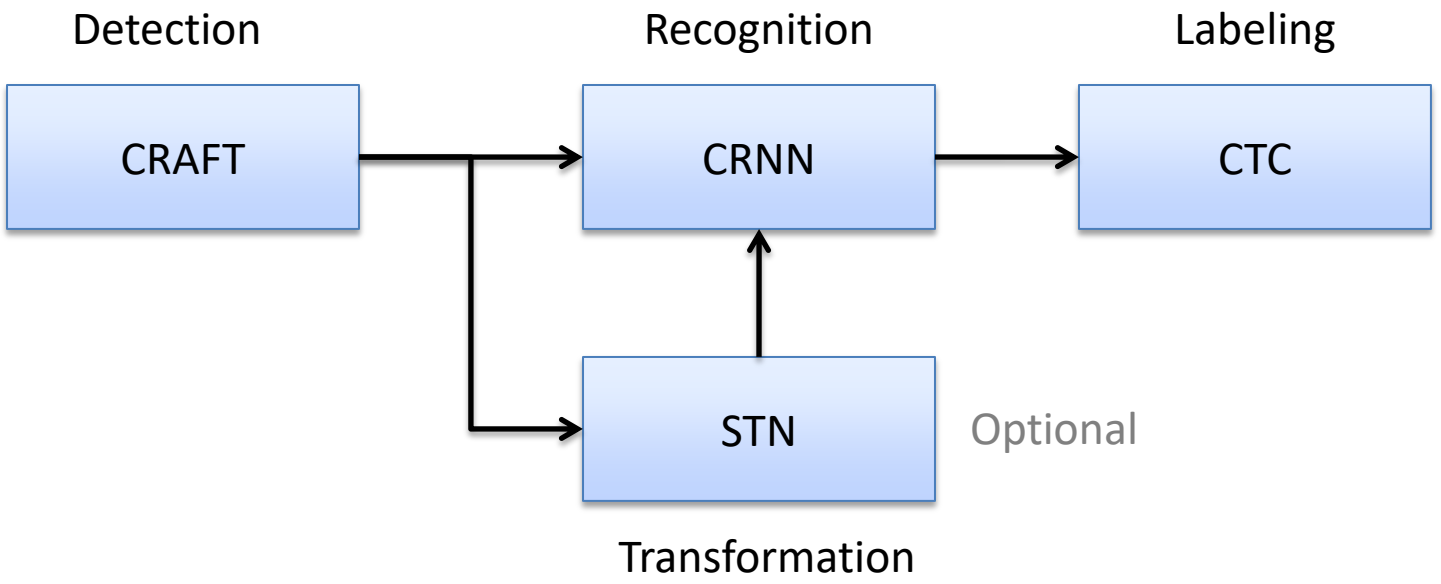
Text Recognition

The second phase of Text Recognition is Text Recognition itself.

Detected Text with Warp Transformation is feed into Text Recognition Model.

Our Text Recognition Model is based on CRNN (Convolutional Recurrent Neural Network)with CTC (Connectionist Temporal Classification).

Scene Text Recognition Pipeline



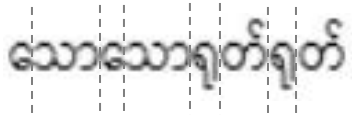
Our Modular Text Recognition has 4 Components: Text Detection Module, Text Transformation Module (Optional), Text Recognition Module and Text Labeling Module.

Burmese Syllables (Word-pieces)

Unlike English Words, Burmese Words are not made up of Alphabets (Characters). Instead, Burmese Words are made up of a group of Glyphs, known as Syllables. Therefore, it is almost impossible to segment a Burmese Word into a group of Characters, which makes it harder for Character Recognition.

To make it worse, the Glyphs of Burmese do not follow the Visual Order; instead they follow the Phonetic Order.

Therefore, in order to solve the Segmentation of Burmese Words, we implement the Visual Order Syllable Segmentation.



ေ + သ + ◌ + သ + ရ + တ် + ရ + တ်

သော + သော + ရုတ် + ရုတ်

ေ + ဖ + င် + ◌ + တ + ီ + ◌

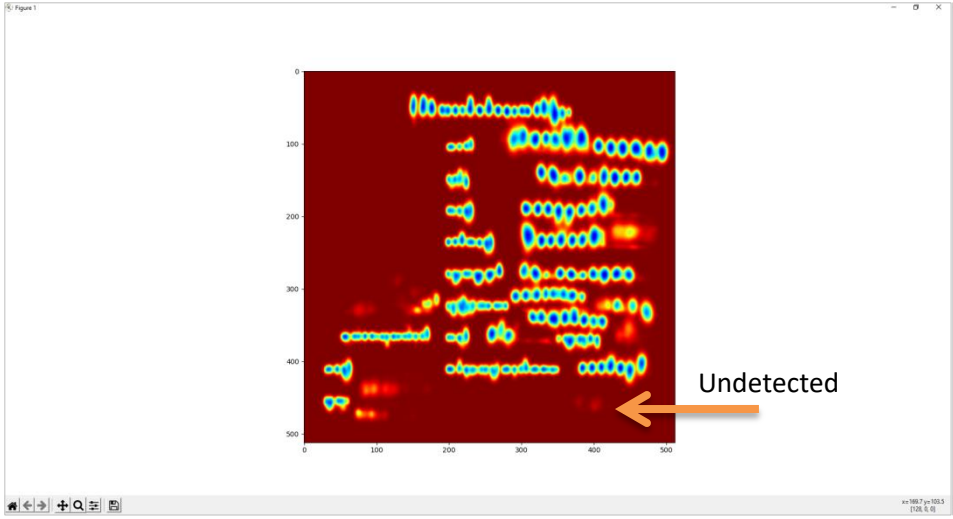
ဖောင် + တော် + ဦး

Generating Scene Text for Detection Dataset

Bilingual Texts (Burmese and English Random Words) are generated with different font styles, colors, orientations and backgrounds, along with Syllables and Bounding Boxes.

Actually, Data Generation is inspired by Synth Text Dataset. However, Synth Text Dataset only has English Characters; therefore, Text Generation for Burmese Characters is essential.

NRC - TEST



Gaussian Heat-map

Text Bounding Boxes

For a Test Case, we will try to detect Myanmar National Registration Card (NRC). As we can see, it is even possible to detect Handwritten Text to some extent. However, to detect Handwritten Text more accurately, further Training Dataset is necessary.

Generating Text for Recognition Dataset

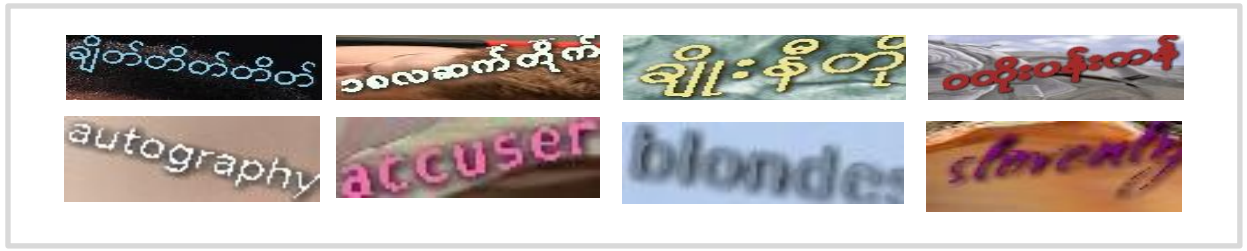
Our Text Recognition Model is based on CRNN Model.

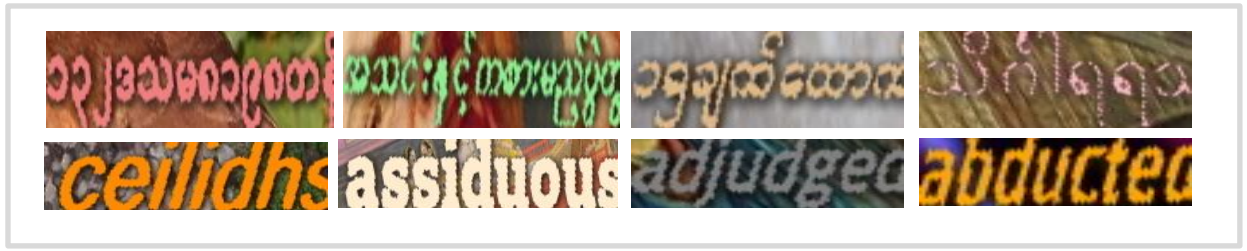
Generating Text Images for Text Recognition involves various steps. Each Text Image is a 128 x 30 Pixel RGB Image in JPEG Format.

Simple Texts, without much augmentations – mostly grayscale, are generated first for Initial Training. Then, more complex Texts, with various augmentations and backgrounds, are generated. Finally, Texts, with Spatial Distortions and Text Decorations, are generated.

In total, there are around 30 Million Images for Recognition Dataset.

Generating Text for Recognition Dataset





SCENE TEXT RECOGNITION - TEST

The following test is taken after 75% of the First Epoch of 15 Million Images, which is current at 4.1% Loss. We will aim to reach less than 0.8% Loss.