COMPUTATIONAL MODELS FOR LUGANDA TEXT RECOGNITION AND TRANSLATION FROM SIGN IMAGES

ABSTRACT

A sign suggests the presence of a fact, condition, or quality that can guide user decisions and actions in helpful ways. A sign is more useful if a person understands its text language. For persons who cannot understand the language used in a sign, there is need for an easily accessible resource they can use to translate to a target language they understand. Such a resource should be capable of capturing an image of the sign, recognizing the text in the image, and translating the recognized text into a convenient target language. In Uganda, there is a growing use of Luganda text in signs. However, Luganda is just one of the relatively under-resourced indigenous Ugandan languages with limited computational resources for text recognition and translation. To contribute in this regard, we adopt a pre-trained convolutional neural network called 'Efficient and Accurate Scene Text (EAST) detector to recognize Luganda text in sign images; thereafter, we apply a Rule-based Machine Translation (RBMT) method to translate recognized text from Luganda to English. The resulting EAST model for Luganda text recognition achieves a precision of 83.27% and a recall of 78.33%. Because of the limited vocabulary of Luganda text in the signs studied so far, the use of a RBMT method ensures perfect translations from the recognized Luganda text to target English text.

Index Terms— Luganda signs, Luganda text detection, Luganda text recognition, Luganda-to-English text translation, Efficient and accurate scene text detector, Rule-based machine translation.

2. RELATED WORKS

2.1 Text detection from images

In general, the methods for detecting text can be categorized in three groups: based on Connected Components (CCs), based on texture, and based on corner.

The methods based on connected components consist of two steps. The first step is to draw CCs from images using a specific method and the second step is to estimate whether the CC is text CC or not based on CCs feature and relative features [1]. The methods for drawing CCs from images can be categorized in three groups: based on edges, based on color and based on a combination of edges and color [4]. The methods based on texture are text/non-text classification methods which deal with text regions as a special texture. A region is identified as a text region or not according to the extracted relevant texture of the candidate regions. In this paper, we use a hybrid approach (which takes the advantages of both CC-based methods and texture-based methods to robustly detect and localize texts in natural scene images. The methods based on corner are inspired by the observation that the character, especially in the text and the caption, usually contains multiple corner points. The aim is to describe the text regions formed by the corner points using several discriminative features [6]. For each text region, text extraction is done to separate text pixels from their backgrounds. In this paper, we pixels in an image are defined as random variables in a Markov Random Field (MRF). To determine the quality of banalization, we use a new energy (or cost) function (introduced in [7]) on these variables, each of which takes a foreground or background label.

2.2 Machine translation

Machine translation deals with the translation of source language (SL) sentences into target language (TL) sentences. A lot of work has been done on machine translation in the last 50 years and there are several approaches that have been established along the way including: Direct translation, Rule-based machine translation, and data driven machine translation [3]. Each of these approaches appeals to specific applications and purposes. In this paper, we employ the rule-based machine translation approach as it appropriately suits the end application where we have a manageable vocabulary of Luganda text in signs.

3. METHODOLOGY

This paper aimed to achieve three specific objectives: to collect and prepare an adequate experimental dataset for Luganda text recognition and translation; to learn Luganda text recognition models and develop a RBMT system; and to evaluate Luganda text recognition models for accuracy.

3.1 Data Collection Techniques

Data was collected from Observed, recorded, organized, categorized or defined information. The main data sources include: Buganda tourist sites like Kasubi Tombs and the Buganda Kingdom Palace in Bulange Nagalabi; published Luganda literature including printed newspapers, printed magazines like entaanda ya Buganda and printed books like the recently published Luganda Bible. For machine translation, the first researcher manually transcribed text from Luganda to English. In total, 200 images were collected and prepared.

The images are comprised of colored images and each of them contain articulate Luganda text. A sample of the dataset is shown in Figure 1

Figure 1: A sample of sign images containing Luganda text

3.2 Development of Luganda text recognition models and translation system

A state-of-the-art pre-trained convolutional neural network called 'Efficient and Accurate Scene Text (EAST) detector was adopted to model the recognition of Luganda text from sign images.

The EAST detector model has proved to be an efficient and fast approach towards text recognition. The key component of this model is a neural network that predicts the existence of text instances in their various geometries from full images. The model is a fully convolutional neural network that outputs dense per pixel predictions of words or text lines. This method eliminates the need for intermediate steps such as candidate proposal, text region formation and word partitioning. During post processing, the method only includes thresholding and Non Maxima Suppression (NMS) on the predicted geometric shapes.

The architecture for the EAST model is loosely inspired by Densebox, where an image is fed into a Fully Convolutional Network (FCN) and multiple channels of pixel-level text score map and geometry are generated. Densebox is a unified FCN that directly predicts bounding boxes and object class confidences through all locations and scales of an image. In our context however, one of the predicted channels is a score map whose pixel values are in the range [0,1]. The remaining channels however represent geometries that enclose the word from the view of each pixel.
The score represents the confidence of the geometry shape predicted at the same location. Ideally, scores closer to 1 represents 100% confidence while 0 represents no or zero confidence, therefore a good model should aim to achieve a score as close to 1 as possible.

Figure 2. A sample of scene images with lines drawn around the texts that was extracted.

3.2.2 Fine tuning a pre-trained EAST feature extractor

In the feature extraction stem we used ImageNet weights as an initializer and then the ICDAR dataset as used in the vanilla EAST model.

The feature merging branch was used as provided in the EAST model; no parameters were modified at this stage. We however froze the weights and did not retrain this branch. The output tensors therefore only flowed into the output layer. We modified the output tensor of the model to represent the classes for Luganda text recognition. Since we have an arbitrary number of classes, we use a SoftMax activation function with 26 neurons each representing one of 26 letters of the alphabet.

The EAST model was pre-trained on the ImageNet database, and then fine-tuned on the ICDAR dataset, and lastly fine-tuned on our dataset. We used the Adam optimizer with the learning rate fixed to 1e-4. We augmented the images to increase the variability of our dataset. Augmentation involved strategies such as horizontal flipping, vertical flipping and hue saturation. During post processing, we performed Non Maxima Suppression (NMS) and set the threshold to 0.3. A list of the hyper-parameters that we used for the EAST model are shown in Table 1

Out of the 200 labeled images, 120 (60%) images were used for training (or learning) Luganda text recognition models.

Parameter	Value
Optimizer	Adam
Learning Rate	1e-4
NMS post processing	0.3
Pre-trained weights	ImageNet, ICDAR 2015

Table 1: Hyper parameters used in the EAST model.

3.2.3 Luganda-to-English RBMT

A Luganda word in this study is taken to be a group of characters written using the Latin alphabet. The group of characters is a valid word only if its defined in the dictionary. A group of words are considered to form a sentence. A Luganda word or a sentence recognized from scene images is provided as input to the lookup program and an entire search or look up of a group of words is done in the dictionary. The initial input is taken to be like a single word unit because some groups of words in Luganda like ”wa ggulu” and ”okuwa amagezi” translate into single English words such as ”above” and ”advise” respectively. A corresponding English translation of the input group of words is then displayed. In cases where the group of words is undefined in the dictionary, a split is done and a check for single words after splitting the

Algorithm 1 Luganda word lookup algorithm

Function (a, b)

Input : (Luganda String *LStr* and English-Luganda dictionary *dict*)

Output: (English Translation)

```
if len(LStr) > 1 then
    search(dict, LStr) if LStr in dict() then
        return dict.value()
    else
        newStr = LStr.split()
        Search(newStr, dict)
        get IndividualDictValues
        return (merge for IndividualDictValues):
    end
else
    return “Sorry that Luganda word doesn’t exist in English”:
end
```

Figure 4. The Luganda word lookup and translation algorithm.

input group of text is conducted. Once a translation of the individual word is found its English translation is displayed as output and this is the output for the Luganda sign recognition and translation model. In cases where the number of words after splitting are more than one, a search for each word translation formed from the split is done in different dictionary indexes. The output after multiple search for multiple words in multiple locations is combined as single output to form a single English sentence translation and this is displayed for the person seeking aid in translating Luganda text in natural scene images. The lookup algorithm is as shown in Figure 4.

3.3 Evaluation

A major goal of any text recognition task is to achieve text recognition models that are satisfactorily accurate. So recognition accuracy is very important. The Luganda text recognition models were automatically evaluated on character accuracy (CA) and character error rate (CER).

$$Character\ Accuracy = \frac{N_c}{N_t} \times 100$$

Where: N_c is the number of correctly extracted characters and N_t is the number of total characters.

To establish the robustness of the models, different noisy conditions were simulated including changes in image resolution, camera view angles, and level(s) of lighting. 80 (40%) of the 200 images were used for testing (evaluation).

We also used precision and recall as additional evaluation measures:

$$Precision = \frac{CD}{CD + FP}$$

$$Recall = \frac{CD}{CD + FN}$$

Where: CD - Correctly detected text, FP – Falsely detected as correct and FN – Falsely detected as incorrect.

4. OVERALL TEXT RECOGNITION AND TRANSLATION SYSTEM DESIGN

In order to achieve flexibility, modularization of the prototype system into three modules: capture module, interactive module, and recognition and translation module has been done. These modules work in client/server mode and are independent of each other. They can be on the same machine or different machines.

The capture module handles image input and it is hardware dependent. The module supports images for both window and DirectX formats. A picture is entered into the recognition and translation module for processing. The recognition and translation module is a key part of the model. The module

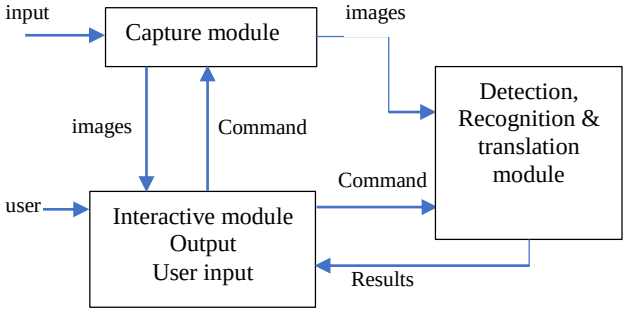

Figure 5 Module interactions.

12. REFERENCES

[1] Yao.C, Bai.X, Sang.N, Zhou.X, and Zhou.S, "Scenetext detection via holistic, multi-channel prediction," pp. 1–10, 2016. 48

[2] D.S.Kim and S.I.Chien, "Automatic Car License Plate Extraction using Modified Generalized Symmetry Transform and Image Warping," Proceedings of International Symposium on Industrial Electronics, Vol. 3, 2001.

[3] T. Dinh, J. Park, and G. Lee, "Low-Complexity Text Extraction in Korean Signboards for Mobile Applications," IEEE International Conferenceon Computer and Information Technology, 2008.

[4] A.N.Lai and G.S.Lee, " Binarization by Local K-means Clustering for Korean Text Extraction," IEEE Symposium on Signal Processing and Information Technology, 2008.

[5] S. Hassanzadeh and H. Pourghassem, "Fast Logo Detection Based on Morphological Features in Document Image," 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, 2011.

[6] W. Fan, J. Sun, Y. Katsuyama, Y. Hotta, and S. Naoi, "Text Detection in Images Based on Grayscale Decomposition and Stroke Extraction," Chinese Conference on Pattern Recognition, IEEE, 2009.

[7] J.Gao and J. Yang, "An Adaptive Algorithm for Text Detection from Natural Scenes," Proceedings of Computer Vision and Pattern Recognition, 2001.

[8] A. Mishra, K. Alahari, and C. Jawahar, "An MRF Model for Binarization of Natural Scene Text," 2011.

[9] X.F.Wang, L.Huang, and C.P.Liu, "A Novel Method for EmbeddedText Segmentation Based on Stroke and Color," 2011.

[10] B. Rim and L. Schiaratura, "Gesture and speech in fundamentals of nonverbal behavior," 01 1991.

[11] E. Coronado, J. Villalobos, B. Bruno, and F. Mastrogiovanni, "Gesture based Robot Control: Design Challenges and Evaluation with Humans," 05 2017.

[12] Xinyu.Zhou, Cong.Yao, He.Wen, Yuzhi.Wang, and Shuchang.Zhou, "East: An efcient and accurate scene text detector," 2017.

[13] Liu.X, Liang.D, Yan.S, Chen.D, and Qiao.Y, "Fots fast oriented text spotting with a unified network," pp. 1–10, 2018. 48

[14] Yao.C, Bai.X, Sang.N, Zhou.X, and Zhou.S, "Scenetext detection via holistic, multi-channel prediction," pp. 1–10, 2016. 48