


Article

# Unconstrained Bilingual Scene Text Reading Using Octave as a Feature Extractor

Direselign Addis Tadesse , Chuan-Ming Liu \*  and Van-Dai Ta 

Department of Computer Science and Information Engineering, National Taipei University of Technology (Taipei Tech), Taipei 106, Taiwan; t106999405@ntut.edu.tw (D.A.T.); t104999002@ntut.edu.tw (V.-D.T.)

\* Correspondence: cmliu@csie.ntut.edu.tw; Tel.: +886-2-27712171 (ext. 4251)

Received: 16 February 2020; Accepted: 21 June 2020; Published: 28 June 2020



**Featured Application:** The potential applications of scene text reading are ordering large pictures and video databases by their literary substance, such as Bing Maps, Apple Maps, and Google Street View, as well as supporting visual impaired people.

**Abstract:** Reading text and unified text detection and recognition from natural images are the most challenging applications in computer vision and document analysis. Previously proposed end-to-end scene text reading methods do not consider the frequency of input images at feature extraction, which slows down the system, requires more memory, and recognizes text inaccurately. In this paper, we proposed an octave convolution (OctConv) feature extractor and a time-restricted attention encoder-decoder module for end-to-end scene text reading. The OctConv can extract features by factorizing the input image based on their frequency. It is a direct replacement of convolutions, orthogonal and complementary, for reducing redundancies and helps to boost the reading text through low memory requirements at a faster speed. In the text reading process, features are first extracted from the input image using Feature Pyramid Network (FPN) with OctConv Residual Network with depth 50 (ResNet50). Then, a Region Proposal Network (RPN) is applied to predict the location of the text area by using extracted features. Finally, a time-restricted attention encoder-decoder module is applied after the Region of Interest (RoI) pooling is performed. A bilingual real and synthetic scene text dataset is prepared for training and testing the proposed model. Additionally, well-known datasets including ICDAR2013, ICDAR2015, and Total Text are used for fine-tuning and evaluating its performance with previously proposed state-of-the-art methods. The proposed model shows promising results on both regular and irregular or curved text detection and reading tasks.

**Keywords:** octave convolution; bilingual scene text reading; Ethiopic script; attention

## 1. Introduction

Currently, reading text from a natural image is one of the hottest research issues in computer vision and document processing. It has many applications including ordering large pictures and video databases by their literary substance, such as Bing Maps, Apple Maps, Google Street View, and so on. Moreover, it allows for image mining, office automation, and support for the visually impaired. Thus, scene text is highly important for thoughtful and uniform services throughout the world. However, reading text from natural images poses several challenges, due to the use of different fonts (color, type, and size) and texts being written on more than one script. Moreover, imperfect image condition causes distorted text, and complex and inference backgrounds cause unpredictability. As a result, reading or spotting texts from a natural image becomes a challenging task.

Previously, several considerable research outputs were presented for scene text detection [1–5] and scene text recognition [6,7] independently, which led to a computational complexity and integration

problem being used as a text-reading task. To improve these, an end-to-end scene text spotting method was presented in references [8–10], but it still needs improvement in terms of recognition accuracy, memory usage, and speed. For instance, in [11,12] a fully conventional network is applied for scene text detection and recognition by considering the detection and recognition problems independently. For scene text detection, a convolutional neural network (CNN) was applied to extract feature maps from the input image, and then different decoders were used to decode and detect the text region based on the extracted features [5,13,14].

Using the extracted sequences of features at the scene text detection phase, characters/words have been predicted with sequence prediction models [15,16]. These types of approaches led to heavy time cost and ignored the correlation in visual cues for images with a number of text regions, whereas both operations had real integrations. In general, previously proposed scene text detection and recognition approaches were problematic, especially when texts in the image are written in more than one script, different text sizes and text shapes are irregular. Furthermore, most research focused on English language and only a few presented other languages such as Arabic and Chinese. Except our previously presented scene text recognition method [17], there is no research output for scene text reading as well as scene text detection for Ethiopic script-based languages. Ethiopic script is used as a writing system for more than 43 languages, including Amharic, Geez, and Tigrigna.

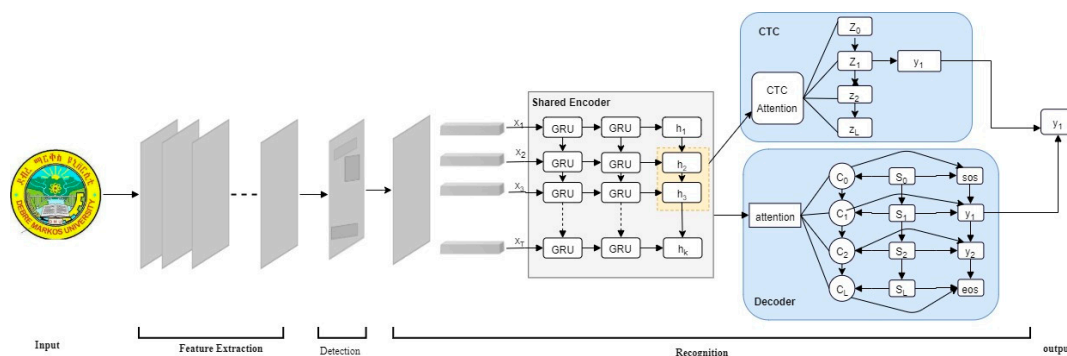
Amharic is the official language of Ethiopia and the second-largest Semitic language after Arabic [18]. On the other hand, English is used as a teaching medium in secondary schools and higher education. As a result, English and Amharic languages are being used concurrently for different activities in most areas of the country. Thus, designing independent applications of scene text detection and scene text recognition requires multiple networks for solving individual sub-problems, which increases computational complexity and causes accuracy and integrity problems. Additionally, developing detection and recognition as independent sub-problems restrains the recognition of rotated and irregular texts. The characteristics of individual characters for complex languages, for example, Amharic language, in the script, and the availability of bilingual scripts in natural images make the scene text recognition methods to challenging when used independently for detection and recognition. Text detection and text recognition are relevant tasks in most operations and complement each other.

Recently, the proposed multilingual end-to-end scene text spotting system in [9,15,19] had a good result for several languages except for Ethiopic script-based languages. However, in their proposed method, they did not consider the frequency of features (high and low) and the effects of word length in the recognition. In this paper, a bilingual end-to-end trainable scene text reading model is proposed by extracting features from the input image based on their frequency and a time-restricted self-attention encoder-decoder module for recognition. Between the feature-extraction and recognition layers, we use a region proposal network, to detect the text area and predict the bounding boxes.

Figure 1 shows the architecture of the proposed system, which contains feature-extraction, detection, and recognition layers. In the first layer of our proposed network, we use a feature pyramid network (FPN) [20] with ResNet-50 to extract features. Inspired by reference [21], the ResNet-50 vanilla convolutions are replaced by octave convolutions (OctConv), except for the first convolution layer. The OctConv factorizes feature tensors based on their frequencies (high and low) which helps to effectively enlarge the receptive field in the original pixel space and improve recognition performance. Additionally, it optimizes the memory requirement by avoiding redundancy. As stated in [21], OctConv improves object-recognition performance and shows a state-of-the-art result. In the second layer, a region proposal network (RPN) is applied for predicting text/non-text regions and recognizing the bounding boxes of the predicted text region from the input image using the extracted feature at the first layer. Finally, by applying Region of Interest (RoI) pooling based on the predicted bounding boxes to the extracted features, word prediction is performed using a time-restricted self-attention encoder-decoder module. Our proposed bilingual text-reading model is originally presented to read texts from the natural image in an end-to-end manner. The major contributions of the article are summarized as follows:

1. Following [22], we prepare large syntactically generated bilingual (English and Amharic) scene text datasets. Additionally, we collect real datasets that have different shapes and written using the two scripts.
2. Our proposed model extracts feature by factorizing based on their frequencies (low and high), which helps to reduce both storage and computation costs. This also helps each layer gain a larger receptive field to capture more contextual information.
3. The proposed system can detect and read texts from an image that has arbitrary shapes, containing oriented, horizontal, and curved text.
4. The performance of the time-restricted attention encoder-decoder module is examined to predict words based on the extracted and segmented features.
5. Using the prepared dataset and well-known datasets, we perform several experiments and our model shows promising results.

The rest of the paper is organized as follows. Related works are presented in Section 2. In Section 3, we discuss the proposed bilingual end-to-end scene text reading methodology. A short description of the Ethiopic script and datasets that are used for training and evaluating the proposed model is described in Section 4. The experimental set-up and results are discussed in Section 5. Finally, a conclusion is drawn in Section 6.



**Figure 1.** The architecture of the proposed bilingual end-to-end scene text reader model.

## 2. Related Work

Reading text from a natural image is currently an active field of investigation in computer vision and document analysis. In this section, we introduce related works, including scene text detection, scene text recognition, and text spotting (combining detection and recognition) techniques.

### 2.1. Scene Text Detection

Traditional and deep-learning machine-learning methods are used to detect texts from a natural image. In [1,3,23–25], scene text detection methods have been presented to detect and bind text areas from a natural image, but this approach has manual computation problems. Lee et al. [25] presented sliding-window-based methods measured by shifting over the image and determining text proximity based on local image highlights. In [26,27], a connected component analysis method was presented to detect scene texts using Stroke Width Transform (SWT) and Maximum Stable Extreme Region (MSER), respectively. However, these approaches are limited when it comes to detecting text regions from distorted images.

Recently, deep-learning techniques improved several machine-learning problems, including scene text detection and recognition problem. Tian et al. [1] presented a Connectionist Text Proposal Network (CTPN), which uses a vertical anchor mechanism that jointly predicts location and text/no-text scores of each fixed width. Shi et al. [14] introduced Segment Linking (SegLink), which is an oriented scene text detection method that segments and then links the text to complete instances using a linkage

prediction. Ma et al. [28] presented a novel rotation-based framework to detect arbitrarily oriented texts found in natural images by proposing region proposal network (RPN) and rotation RoI pooling. A deep direct regression-based method for detecting multi-oriented scene text has been presented in [29]. Efficient and accuracy scene Text detector (EAST) [5] has been introduced to effectively detect words or text lines using a single neural network.

## 2.2. Scene Text Recognition

In the text-reading phases of natural images, text recognition is the second phase after scene text detection. This method can be implemented independently or after scene text detection phases. In the scene text recognition phase, the cropped text regions are fed either from the scene text detection phase or from the prepared input dataset, from which the sequences of labels are decoded. Previous attempts were made by detecting individual characters and refining misclassified characters. Such methods require training a strong character detector for accurately detecting and cropping each character out from the original word. These types of methods are more difficult for Ethiopic scripts due to their complexities. Apart from the character level methods, word recognition [12], sequence to label [30], and sequence to sequence [31] methods have been presented. Liu et al. [32] and Shi et al. [15] presented a spatial attention mechanism to transform a distorted text region from irregular input images into canonical pose suitable recognition. However, both the detection and recognition task performance are determined based on the extracted features. Previously proposed scene text detection and recognition of deep learning-based and conventional machine learning feature extraction methods do not consider the frequency of the input image. Following [21], in this paper, we propose an OctConv with ResNet-50 feature extractor, which extracts features by factorizing based on their frequencies.

## 2.3. Scene Text Spotting

Recently, several end-to-end scene text spotting methods have been introduced and have shown a remarkable result compared to independent scene text detection and recognition approaches. For instance, Li et al. [10] introduced an end-to-end text spotting technique from natural images using RPN as a text detector and attention Long Short Term Memory (LSTM) as a text recognizer. Liao et al. [8] presented an end-to-end scene text-reading method using Single Shot Detector (SSD) [33] and convolutional recurrent neural network (CRNN) for scene text detection and recognition, respectively. Liu et al. [34] introduced a unified network to detect and recognize multi-oriented scene texts from natural images. Lunadren et al. [35] introduced an octave-based fully convolutional neural network with fewer layers and parameters to precisely detect multilingual scene text. The most recently proposed scene text-reading models are summarized in Table 1.

**Table 1.** Summary of recently proposed end-to-end scene text recognition models.

Method	Model	Detection	Recognition	Year
Liao et al. [11]	TextBoxes	SSD-based framework	CRNN	2017
Büsta et al. [19]	Deep TextSpotter	Yolo v2	CTC	2017
Liu et al. [34]	FOTS	EAST with RoI Rotate	CTC	2018
Liao et al. [8]	TextBoxes++	SSD-based framework	CRNN	2018
Liao et al. [9]	Mask TextSpotter	Mask R-CNN	Character segmentation + Spatial attention module	2019

Improving the feature extraction and recognition network will improve scene text detection, recognition, and text spotting problems. In [21], an OctConv feature extraction method has been proposed for object detection and improves its performance. Octave convolution addresses spatial redundancy, which was not addressed in the previously proposed methods. The OctConv does not change the connectivity between feature maps and it is different from inception multi-path designs [36,37]. In our proposed bilingual text-reading method, we replace the ResNet-50 vanilla

convolution with OctConv, which can operate quickly and produce accurate results in the extraction of features. As stated in [38], the limitation of Connectionist Temporal Classification (CTC), attention encoder-decoder, and hybrid (CTC and attention) method is improved using a time-restricted self-attention method for an automatic speech recognition system. In our proposed method, we integrate a time-restricted self-attention encoder-decoder module for recognition with feature extraction and bounding box detection layers.

### 3. Methodology

In this section, the details of the proposed bilingual scene text-reading model are presented. The architecture of the model, shown in Figure 1, is trained in an end-to-end manner that concurrently detects and recognizes words from a natural image.

#### 3.1. Overall View of the Architecture

Our proposed architecture follows the architecture presented in [9,21]. Our proposed architecture has three functional components, feature-extraction layer, text/non-text detection layer, and recognition layer. In the feature-extraction layer, features are extracted from input natural images and passed to the next layer using an FPN [20] with ResNet-50 [39] by replacing the vanilla convolution with an octave convolution. Then, using the extracted features on the 1st layer as an input, a region proposal network (RPN) [40] predicts text/non-text area and bounding boxes of each text area. Finally, by applying RoI to the outputs of the 2nd layer, text segmentation, and word prediction are done using the time-restricted self-attention encoder-decoder module. Details of each layer are presented below.

#### 3.2. Feature Extraction Layer

Feature extraction is one of the crucial steps in machine learning problems. In the deep learning era, several automatic feature extraction methods have been proposed, including [40–43]. These feature extraction methods were applied to several problem domains and produced good results. Recently, Chen et al. [21], proposed an OctConv method that extracts features based on their frequencies. We use Chen et al.'s feature extraction method to detect text/non-text regions. Naturally, texts found in natural images have different properties (i.e., size, orientation, shapes, and color). These cause a challenge in perfectly detecting the text/non-text region, which directly affects the performance of the recognition task. To overcome this challenge, we build high-level semantic feature maps using FPN with ResNet-50. Different from [9], in our proposed feature extraction layer, we replace vanilla convolutions by OctConv. This factorizes the mixed-feature map tensor into high and low-frequency maps, where the high-frequency feature map tensors encode with fine details, whereas the low-frequency feature map tensors encode with global structures. Compared to vanilla convolution, OctConv reduces spatial redundancy, memory cost, and computation cost.

For given spatial dimensions  $w$  and  $h$  with the number of feature maps  $c$ , the input feature tensor of a convolution layer will be  $X \in \mathbb{R}^{c \times h \times w}$ . In OctConv, the input vector  $X$  factorized along channel dimensions into low feature map ( $X^L$ ) and high feature map ( $X^H$ ) frequencies. As stated in [21], the factorization of high feature map and low feature map tensors are computed as follows:

$$X^H = X^{(1-\alpha)c \times h \times w} \quad (1)$$

$$X^L = X^{\alpha c \times \frac{h}{2} \times \frac{w}{2}} \quad (2)$$

where the value of  $\alpha \in [0, 1)$

In the factorization process, fine details are obtained on high-frequency feature maps, whereas differences in speed in spatial dimensions with respect to image location were obtained at low-frequency feature map tensors. This process maps the features that are compacted and replace spatial repetitive feature maps with different resolution maps. On these feature maps, an octave convolution is applied



where the vanilla convolution does not work, due to different resolutions of high- and low-frequency feature maps. The octave convolution enables efficient inter-frequency communication and effectively operates on low- and high-frequency tensors. For the factorized high ( $X^H$ ) and low ( $X^L$ ) feature tensors, there is a corresponding output feature tensor  $Y^H$  and  $Y^L$ , respectively. To get each output feature tensor, inter ( $Y^{H \rightarrow L}, Y^{L \rightarrow H}$ ) and intra ( $Y^{L \rightarrow L}, Y^{H \rightarrow H}$ ) frequency convolution update is performed. Each output feature map at location  $(p, q)$  is computed using appropriate kernels ( $W^L$  and  $W^H$ ), applying regular convolution for intra-frequency update and removing the need of explicitly computing and sorting on up/down sampling for inter-frequency communication as follows:

$$Y_{p,q}^H = \sum_{i,j \in N_k} \left( W_{i+\frac{k-1}{2}, j+\frac{k-1}{2}}^{H \rightarrow H} \right)^T X_{p+i, q+j}^H + \sum_{i,j \in N_k} \left( W_{i+\frac{k-1}{2}, j+\frac{k-1}{2}}^{L \rightarrow H} \right)^T X_{(\frac{p}{2}+i), (\frac{q}{2}+j)}^L \quad (3)$$

$$Y_{p,q}^L = \sum_{i,j \in N_k} \left( W_{i+\frac{k-1}{2}, j+\frac{k-1}{2}}^{L \rightarrow L} \right)^T X_{p+i, q+j}^L + \sum_{i,j \in N_k} \left( W_{i+\frac{k-1}{2}, j+\frac{k-1}{2}}^{H \rightarrow L} \right)^T X_{(2^*p+0.5+i), (2^*p+0.5+j)}^H \quad (4)$$

The recognition performance of the model is improved because OctConv can extract a larger receptive field for low-frequency feature maps. Most commonly, text found in natural images has low frequencies. Compared to vanilla convolution, OctConv convolves at a factor of 2 receptive fields.

### 3.3. Text Region Detection Layer

Using RPN and taking the extracted feature maps as an input, text/non-text regions are detected. Following [9] and [20], we assign five anchors at different stages {P2, P3, P4, P5, P6} with the area of anchors  $\{32^2, 64^2, 128^2, 256^2, 512^2\}$ , respectively. Besides, to handle different text sizes  $\{0.5, 1, 2\}$  aspect ratios are implemented at each stage. By doing this, text proposal features are generated. These features are further extracted using RoI align [41], which preserves a more accurate location compared to RoI pooling. Finally, the Fast Region (R)-CNN [41] generates precise bounding boxes for the texts found in the input natural image. Using a soft-Non-maximal suppression (NMS) [42] technique, we select one bounding box for those texts that have more than one bounding box.

### 3.4. Segmentation and Recognition Layer

After texts are detected at the detection layer, text segmentation and recognition of words are performed. Text instance regions are segmented using four consecutive convolution layers with  $3 \times 3$  filters and deconvolution layers with  $2 \times 2$  filters and strides on the outputs of RoI align feature in the previous layer, with predicted bounding boxes. Finally, the outputs of the segmented text instance feature  $x = (x_1, x_2, \dots, x_T)$  are fed for a time-restricted self-attention encoder-decoder module.

In [43], a time-restricted (attention window) self-attention encoder-decoder module is presented for automatic speech recognition, which produces a state-of-the-art result by improving the limitations of CTC (i.e., hard alignment problem and conditional independence constraints) and the attention encoder-decoder module. Unlike [9], we use a time-restricted self-attention module using a bidirectional Gated Recurrent Unit (GRU) as an encoder and a GRU as a decoder. Form the extracted and segmented features, the bidirectional encoder computes the hidden feature vector  $h_t$  as follows:

$$z_t = \sigma(W_{xz}x_t + U_{hz}h_{t-1} + b_z) \quad (5)$$

$$r_t = \sigma(W_{xr}x_t + U_{hr}h_{t-1} + b_r) \quad (6)$$

$$h_t = \tanh(W_{xh}x_t + U_{rh}(r_t \otimes h_{t-1}) + b_h) \quad (7)$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes h_t \quad (8)$$

where  $z_t$ ,  $r_t$ ,  $h_t$ , and  $h_t$  are update gate, reset gate, current memory, and final memory at the current time step, respectively.  $W$ ,  $U$ , and  $b$  are parameter matrices and vector;  $\sigma$  and  $\tanh$  stand for sigmoid and hyperbolic tangent function, respectively.

Using the embedding matrix  $W_{emb}$  the hidden vector  $h_t$  is converted to embedding matrix  $b_t$  as follows:

$$b_t = W_{emb}h_t, t = u - \tau, \dots, u + \tau \tag{9}$$

By applying a linear projection on the embedded vector  $b_t$  query ( $q_t$ ), values ( $v_t$ ), and keys ( $k_t$ ) vectors are computed as follows:

$$q_t = Qb_t, t = u \tag{10}$$

$$k_t = Kb_t, t = u - \tau, \dots, u + \tau \tag{11}$$

$$v_t = Vb_t, t = u - \tau, \dots, u + \tau \tag{12}$$

where  $Q$ ,  $K$ , and  $V$  are query, key and value matrices, respectively.

Based on these results, attention weight  $a_u$  and attention result  $c_u$  are derived as follows:

$$e_{ut} = \frac{q_u^T k_t}{\sqrt{d_k}} \tag{13}$$

$$a_{ut} = \frac{\exp(e_{ut})}{\sum_{t'=1}^{u+\tau} \exp(e_{ut'})} \tag{14}$$

$$c_u = \sum_{t=u-\tau}^{u+\tau} a_{ut} h_t \tag{15}$$

To address the conditional independence assumption in CTC, an attention layer is placed before the CTC projection layer  $ph_u$  and transforms it to a particular dimension representing the number of CTC output labels. Then, the attention layer output that carries context information is served as the input of CTC projection layer at the current time  $u$ .

$$ph_u = W_{proj}c_u + b \tag{16}$$

where  $W_{proj}$  and  $b$  are the weight matrix and bias of the CTC projection layer, respectively.

Finally, the projected output is optimized as follows:

$$L_{CTC} = -\log \sum_{\pi \in B^{-1}(y)} p(\pi|ph_u) \tag{17}$$

where  $y$  denotes the output label sequence. A many-to-one mapping  $B$  is defined to determine the correspondence between a set of paths and the output label sequences. The self-attention layer links all positions with a constant number of operations that are performed in sequence.

#### 4. Ethiopian Script and Dataset Collection

##### 4.1. Ethiopian Script

Ethiopic script, which is derived from Geez, is one of the most ancient scripts in the world. It is used as a writing system for more than 43 languages, including Amharic, Geez, and Tigrigna. The script has largely been used by Geez and Amharic, which are the liturgical and official languages of Ethiopia, respectively. Amharic language is the second Semitic language after Arabic. The script is written down in a tabular format in which the first column denotes the base character and the other columns are vowels derived from the base characters, made by slightly deforming or modifying the

base characters. The script has a total of 466 characters, out of which 20 are digits, 9 are punctuation marks, and the remaining 437 characters are parts of the alphabet. Developing a scene text recognition system for Ethiopic script is challenging, due to the visually similar characters, especially between base characters and the derived vowels, and the number of characters in the script. Furthermore, the lack of training and testing datasets is another limitation in the development of a scene text reading system for Ethiopic scripts. In this paper, we propose an end-to-end trainable bilingual scene text reading model using FPN, RPN, and time-restricted self-attention CTC.

#### 4.2. Dataset Collection

In any machine learning technique, a dataset plays an important role in training and obtaining a better machine learning model. In particular, deep learning methods are more data-hungry than traditional machine learning algorithms. However, preparing a large dataset was a challenging task specifically for under-resourced languages. In this paper, we use a syntactically generated scene text dataset, and real scene text dataset for training and testing the proposed model, respectively. Following [12], a bilingual scene text dataset is prepared. A detailed description of synthetic dataset generation and real scene text dataset preparation is provided in the following sections.

##### 4.2.1. Synthetic Scene Text Dataset

To train the proposed model, we use a bilingual scene text dataset, which is generated by adding a simple modification to the scene text dataset generation technique presented in [12]. The generated scene text images are like real scene images. This technique is very important to get more training data for those scripts that do not have prepared real scene text datasets. As far as we know, there is no prepared real scene text dataset for Ethiopic script. Moreover, most texts found in natural images are written in two languages (Amharic and English). Due to this, we prepare 500,000 bilingual training datasets from 54,735 words (825,080 characters), which were collected from social, political, and governmental websites that are written in Amharic and English. In the dataset generation process, 72 freely available Ethiopic Unicode fonts, different background images, font size, rotation along the horizontal line, and skew and thickness parameters are tuned. The sample generated scene image and statistics of the generated dataset are presented in Figure 2 and Table 2, respectively.

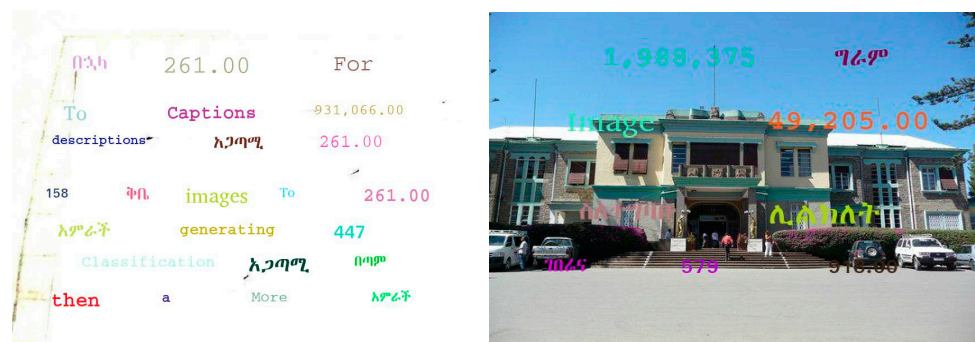


Figure 2. Sample of synthetically generated scene text images.

##### 4.2.2. Real Scene Text Dataset

In addition to the synthetic dataset, we collected 1200 benchmark bilingual real scene text images using photo camera and image search on Google. The images were captured from local markets, navigation and traffic signs, banners, billboards, and governmental offices. We also incorporated several office logos, most of which were written both in Amharic and English with curved shapes. In addition to our prepared dataset, we used the Synthetic [22] dataset to pre-train the proposed model with our synthetic dataset. To refine the pre-trained model and compare its performance with a state-of-the-art model, we used ICDAR2013 [44], ICDAR2015 [40], and Total-Text [45] datasets.



The datasets, we used in the proposed model are summarized in Table 2. Additionally, sample images from the collected datasets are depicted in Figure 3.



Figure 3. Sample of collected real scene text images.

Table 2. Statistics of datasets applied for training and testing the proposed model.

Dataset	Language	Total Images	Training	Testing	Type	
Ours	Real	Bilingual	1200	600	600	Irregular
	Synthetic	Bilingual	500,000	500,000	-	Regular
ICDAR2013 [44]	English	462	229	233	Regular	
ICDAR2015 [40]	English	1500	1000	500	Regular	
Synthetic [22]	English	600,000	-	-	Regular	
Total-Text [45]	English	1555	1255	300	Irregular	

## 5. Experiments and Discussions

The effectiveness of the proposed model was evaluated and compared with state-of-the-art methods by pre-training the proposed model using our synthetically generated dataset and a Synthetic dataset. Finally, the pre-trained model was refined by merging the above-mentioned datasets.

### 5.1. Implementation Details

The proposed model was first pre-trained using our synthetically generated bilingual dataset and Synthetic [22], then fine-tuned using the union of other real-world datasets indicated in Section 4.2.2. Due to the lack of real sample images in the fine-tuning stage, data augmentation and multi-scale training were applied by randomly modifying brightness, hue, contrast, the angle of the image between  $-30$  and  $30$ . Following [9], for multi-scale training, the shorter sides of the input images were randomly resized to five scales (600, 800, 1000, 1200, 1400). We used Adam [46] (base learning rate = 0.0001,  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , weight decay = 0) as an optimizer. Following the result of [21], we set the value to  $\alpha = 0.25$  which denotes the ratio of the low-frequency part.

The experiment of the proposed bilingual scene text reading model is conducted on the Ubuntu machine containing Intel Core i7-7700 (3.60 GHz) CPU with 64 GB RAM and GeForce GTX 1080 Ti 11176 MiB GPU. For the implementation, we use Python 3.7 and PyTorch1.2.

## 5.2. Experiment Results

Throughout our experimental analysis, we evaluated a single model trained in a multilingual setup as explained in Section 3. To improve the performance of the model, we first pre-trained it using Synthetic dataset [22] and our synthetically generated bilingual dataset which has a total of 430 characters. Then, we fine-tuned the pre-trained model by combining the above-mentioned real scene text datasets. The text recognition results were reported in an unconstrained setup, that is, without using any predefined lexicon (set of words).

The performance of the trained model was verified using our prepared testing dataset and well-known ICDAR detests. As discussed in Section 4.2, the collected images in our dataset contain horizontal, arbitrary, and curved texts. Both the detection and recognition results were promising for horizontal, arbitrary, and curved text. The experiment evaluation for scene text detection on our prepared real scene text dataset showed 88.3% Precision (P), 82.4% Recall (R), and 85.25% F1-score (F). On the other hand, the end-to-end scene text-reading experiment showed 80.88% P, 49.01% R, and 61.04% F. The scene text detection performance of the proposed method for English and Amharic words do not differ much. However, in the end-to-end scene text reading task, 63.4% of errors occurred in the recognition of Amharic words. From incorrectly recognized characters, some of them did not have sufficient samples on the real and Synthetic datasets. Sample detection and recognition results are depicted in Figure 4. Most of the detection errors in our proposed method occurred from false detection of non-text areas of backgrounds.



**Figure 4.** Sample detection and recognition result for our prepared dataset.

In addition to our testing dataset, we evaluated the performance of our proposed model using ICDAR2013, ICDAR2015, and Total-Text testing datasets, which contain only English texts. The model is fine-tuned for both English and Amharic languages as one model, not for each language. The results of our proposed method and previously proposed methods are shown in Table 3. The experiment showed that our proposed method had a better recognition result on ICDAR2013 and Total-Text datasets. However, the scene text detection result of our proposed method was almost similar to a recently proposed mask text spotter [9] method. We used their architecture and implementation code with a little modification on the feature extraction layer and recognition layer. From the MaskTextSpotter implementation, we modified the ResNet-50 feature extraction by octave based ResNet-50 feature extraction and the text recognition part is modified by self-attention encoder-decoder model. Whereas the preprocessing and RPN implementation is taken from MaskTextSpotter. In Table 4, we compare the scene text detection result of our proposed method with previously proposed methods using ICDAR2013, ICDAR2015, and Total-Text datasets.

**Table 3.** F1-Score experimental results of the proposed unconstrained scene text reading system compared with previous methods.

Method	ICDAR2013	ICDAR2015	Total-Text
TextProposals+DicNet * [47]	68.54%	47.18%	-
DeepTextSpotter * [19]	77.0%	47.0%	-
FOTS * [34]	84.77%	65.33%	-
TextBoxes * [8]	84.65%	51.9%	-
E2E-MLT ** [48]	-	71.4%	-
Mask Text Spotter ** [9]	86.5%	62.4%	65.3%
Ours	86.8%	62.15%	67.6%

\* indicates that the model is trained for English language only; \*\* indicates that the model is trained for multilingual datasets. Our model is trained for English and Amharic languages, with 430 characters.

**Table 4.** Scene text detection result of the proposed method compared with previous methods.

Method	ICDAR2013			ICDAR2015			Total-Text		
	P	R	F	P	R	F	P	R	F
PSENet [49]	94%	90%	92%	86.2%	84.5%	85.69%	84%	77.9%	80.9%
TextBoxes++ [8]	92%	86%	89%	87.8%	78.5%	82.9%	-	-	-
Mask Text Spotter [9]	94.8%	89.5%	92.1%	86.8%	81.2%	83.4%	81.8%	75.4%	78.5%
Ours	93.91%	88.96%	91.36%	86.02%	80.97%	83.28%	82.3%	73.8%	77.82%

In the experiment, the proposed bilingual scene text reading method had limitations regarding small font size scene texts and severely distorted images. Furthermore, due to the existence of many characters and their similarities, and the limited number of training samples for certain Ethiopic characters, a recognition error occurred at the time of testing. To improve the recognition performance of the system and the scene text-reading system in general, it is necessary to prepare more training data that contain enough samples for every character.

## 6. Conclusions

This paper introduced an end-to-end trainable bilingual (English and Ethiopic) scene text reading system using octave convolution and time-restricted attention encoder-decoder module. In the proposed model there were three layers. In the first layer, FPN with ResNet-50 was used as a feature extractor by replacing vanilla convolution with OctConv. Secondly, bounding box prediction and detection of texts were performed using RPN. Finally, recognition of text was performed by segmenting text areas based on the detected bounding boxes on the second layer using a time-restricted attention encoder-decoder network. To measure the effectiveness of the proposed model, we collect and syntactically generate a bilingual dataset. Additionally, we use well-known ICDAR2013, ICDAR2015, and Total Text datasets. Based on the prepared bilingual dataset, the proposed method shows 61.04% and 85.25% F1-measure on scene text reading and scene text detection, respectively. Compared to state-of-the-art recognition performance, our proposed model shows promising results. However, our method shows state-of-the-art results for ICDAR2013 and Total-Text end-to-end text readings. Furthermore, due to the existence of many characters, their similarities, and the limited number of training samples for certain Ethiopic characters, a recognition error occurred at the time of testing. To improve the recognition performance of the system, it is necessary for the future to prepare more training data that contain enough samples for every character. After the publication of the paper, the implementation code and the prepared dataset link will be freely available for the researchers on [https://github.com/direselign/amh\\_eng](https://github.com/direselign/amh_eng).

**Author Contributions:** Conceptualization D.A.T. and experiments, D.A.T. and V.-D.T.; validation, D.A.T. and V.-D.T.; writing – original draft preparation, D.A.T.; writing-review and editing, C.M.L. and V.-D.T.; super vision, C.-M.L.; funding acquisition, C.-M.L.; All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by the Ministry of Science and Technology and National Taipei University of Technology through the Applied Computing Research Laboratory, Taiwan, under Grant MOST 107-2221-E-027-099-MY2 and Grant NTUT-BIT-109-03.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Tian, Z.; Huang, W.; He, T.; Qiao, Y. Detecting Text in Natural Image with Connectionist Text Proposal Network. Available online: <http://textdet.com/> (accessed on 14 December 2019).
2. Yao, C.; Bai, X.; Sang, N.; Zhou, X.; Zhou, S.; Cao, Z. Scene text detection via holistic, multi-channel prediction. *arXiv* **2016**, arXiv:1606.09002. Available online: <http://arxiv.org/abs/1606.09002> (accessed on 31 March 2019).
3. Buta, M.; Neumann, L.; Matas, J. FASText: Efficient unconstrained scene text detector. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7 December 2015; Volume 2015 Inter, pp. 1206–1214.
4. Deng, D.; Liu, H.; Li, X.; Cai, D. PixelLink: Detecting scene text via instance segmentation. *arXiv* **2018**, arXiv:1801.01315. Available online: <http://arxiv.org/abs/1801.01315> (accessed on 10 February 2019).
5. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. EAST: An efficient and accurate scene text detector. *arXiv* **2017**, arXiv:1704.03155.
6. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Reading text in the wild with convolutional neural networks. *Int. J. Comput. Vis.* **2016**, *116*, 1–20. [[CrossRef](#)]
7. He, P.; Huang, W.; Qiao, Y.; Loy, C.C.; Tang, X. Reading scene text in deep convolutional sequences. *arXiv* **2015**, arXiv:1506.04395. Available online: <https://arxiv.org/abs/1506.04395> (accessed on 2 April 2019).
8. Liao, M.; Shi, B.; Bai, X. TextBoxes++: A single-shot oriented scene text detector. *arXiv* **2018**, arXiv:1801.02765. [[CrossRef](#)] [[PubMed](#)]
9. Liao, M.; Lyu, P.; He, M.; Yao, C.; Wu, W.; Bai, X. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In Proceedings of the 2019 Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; LNCS; Volume 11218, pp. 71–88. [[CrossRef](#)]
10. Li, H.; Wang, P.; Shen, C. Towards end-to-end text spotting with convolutional recurrent neural networks. *arXiv* **2017**, arXiv:1707.03985. Available online: <http://arxiv.org/abs/1707.03985> (accessed on 2 April 2019).
11. Liao, M.; Shi, B.; Bai, X.; Wang, X.; Liu, W. TextBoxes: A Fast text detector with a single deep neural network. *arXiv* **2016**, arXiv:1611.06779. Available online: <http://arxiv.org/abs/1611.06779> (accessed on 2 April 2019).
12. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv* **2014**, arXiv:1406.2227. Available online: <http://arxiv.org/abs/1406.2227> (accessed on 2 March 2019).
13. Tian, S.; Bhattacharya, U.; Lu, S.; Su, B.; Wang, Q.; Wei, X.; Lu, Y.; Tan, C.L. Multilingual scene character recognition with co-occurrence of histogram of oriented gradients. *Pattern Recognit.* **2016**, *51*, 125–134. [[CrossRef](#)]
14. Shi, B.; Bai, X.; Belongie, S. Detecting oriented text in natural images by linking segments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 2550–2558. Available online: [http://openaccess.thecvf.com/content\\_cvpr\\_2017/html/Shi\\_Detecting\\_Oriented\\_Text\\_CVPR\\_2017\\_paper.html](http://openaccess.thecvf.com/content_cvpr_2017/html/Shi_Detecting_Oriented_Text_CVPR_2017_paper.html) (accessed on 11 April 2019).
15. Shi, B.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Robust scene text recognition with automatic rectification. *arXiv* **2016**, arXiv:1603.03915. Available online: <http://arxiv.org/abs/1603.03915> (accessed on 21 March 2019).
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [[CrossRef](#)]
17. Addis, D.; Liu, C.-M.; Ta, V.-D. *Ethiopic Natural Scene Text Recognition Using Deep Learning Approaches*; Springer: Cham, The Netherlands, 2020; pp. 502–511.
18. Simons, G.F.; Fennig, C.D. *Ethnologue: Languages of the World*, 20th ed.; SIL International: Dallas, TX, USA, 2017.



19. Busta, M.; Neumann, L.; Matas, J. Deep Textspotter: An End-to-End Trainable Scene Text Localization and Recognition Framework. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2204–2212.
20. Lin, T.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
21. Chen, Y.; Fan, H.; Xu, B.; Yan, Z.; Kalantidis, Y.; Rohrbach, M.; Yan, S.; Feng, J. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. *arXiv* **2019**, arXiv:1904.05049. Available online: <http://arxiv.org/abs/1904.05049> (accessed on 20 September 2019).
22. Gupta, A.; Vedaldi, A.; Zisserman, A. Synthetic data for text localisation in natural images. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2016; Volume 2016-Decem, pp. 2315–2324.
23. Neumann, L.; Matas, J. Scene Text Localization and Recognition with oriented Stroke Detection. In Proceedings of the IEEE International Conference on Computer Vision 2013, Sydney, Australia, 1–8 December 2013; pp. 97–104.
24. Wang, K.; Babenko, B.; Belongie, S. End-to-end scene text recognition. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1457–1464.
25. Lee, J.J.; Lee, P.H.; Lee, S.W.; Yuille, A.; Koch, C. AdaBoost for text detection in natural scene. In Proceedings of the IEEE International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 429–434.
26. Epshtein, B.; Ofek, E.; Wexler, Y. Detecting text in natural scenes with stroke width transform. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 2963–2970.
27. Matas, J.; Chum, O.; Urban, M.; Pajdla, T. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis. Comput.* **2004**, *22*, 761–767. [[CrossRef](#)]
28. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [[CrossRef](#)]
29. He, D.; Yang, X.; Liang, C.; Zhou, Z.; Ororbial, A.G.; Kifer, D.; Giles, C.L. Multi-scale FCN with cascaded instance aware segmentation for arbitrary oriented word spotting in the wild. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 474–483.
30. Su, B.; Lu, S. Accurate Scene Text Recognition Based on Recurrent Neural Network. In *Computer Vision—ACCV 2014, Lecture Notes in Computer Science*; Cremers, D., Reid, I., Saito, H., Yang, M.H., Eds.; Springer: Cham, The Netherlands, 2014; Volume 9003, pp. 35–48.
31. Lee, C.-Y.; Osindero, S. Recursive recurrent nets with attention modeling for OCR in the wild. Proceeding of the IEEE conference on Computer Vision and Patter Recognition (CVPR 2016), Las Vegas, NV, USA, 26–30 June 2016; pp. 2231–2239.
32. Liu, W.; Chen, C.; Wong, K.-Y.; Su, Z.; Han, J. STAR-Net: A Spatial attention residue network for scene text recognition. *BMVC* **2016**, *2*, 7.
33. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision 2016, Amsterdam, The Netherlands, 8–16 October 2016; pp. 21–37.
34. Liu, X.; Liang, D.; Yan, S.; Chen, D.; Qiao, Y.; Yan, J. Fots: Fast Oriented Text Spotting with a Unified Network. Proceeding of the IEEE conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5676–5685.
35. Lundgren, A.; Castro, D.; Lima, E.; Bezerra, B. OctShuffleMLT: A compact octave based neural network for end-to-end multilingual text detection and recognition. In Proceedings of the 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), Sydney, Australia, 22–25 September 2019; pp. 37–42.
36. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In Proceedings of the 31st AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 4278–4284.

37. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; Volume 07-12-June, pp. 1–9.
38. Povey, D.; Hadian, H.; Ghahremani, P.; Li, K.; Khudanpur, S. A time-restricted self-attention layer for ASR. Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; Volume 2018-April, pp. 5874–5878.
39. Zagoruyko, S.; Komodakis, N. Wide residual networks. *arXiv* **2016**, arXiv:1605.07146.
40. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing System 2015, Montreal, PQ, Canada, 7–12 December 2015; pp. 91–99.
41. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
42. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS—Improving Object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision 2017, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
43. Wu, L.; Li, T.; Wang, L.; Yan, Y. Improving hybrid CTC/attention architecture with time-restricted self-attention CTC for end-to-end speech recognition. *Appl. Sci.* **2019**, *9*, 4639. [[CrossRef](#)]
44. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Bigorda, L.G.i.; Mestre, S.R.; Mas, J.; Mota, D.F.; Almazan, J.A.; de Las Heras, L.P. ICDAR 2013 robust reading competition. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1484–1493.
45. Ch’Ng, C.K.; Chan, C.S. Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition. In Proceedings of the International Conference on Document Analysis and Recognition, Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 935–942.
46. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
47. Gómez, L.; Karatzas, D. TextProposals: A text-specific selective search algorithm for word spotting in the wild. *Pattern Recognit.* **2017**, *70*, 60–74. [[CrossRef](#)]
48. Bušta, M.; Patel, Y.; Matas, J. E2E-MLT—An unconstrained end-to-end method for multi-language scene text. *arXiv* **2018**, arXiv:1801.09919. Available online: <http://arxiv.org/abs/1801.09919> (accessed on 11 April 2019).
49. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape robust text detection with progressive scale expansion network. *arXiv* **2019**, arXiv:1903.12473. Available online: <http://arxiv.org/abs/1903.12473> (accessed on 6 January 2019).



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).