Interpreting Texts in Natural Scene Images

Ong Yi Ling, Lau Bee Theng, Chai Wei Yen Almon, and McCarthy Christopher

***Abstract*—Scene texts in natural images are important as they contain information that helps us understand the scenes. Mobile scene text detection applications have become popular in recent years due to their increasing fast performance and low price availability. Therefore, the 'Let Me Read for You' mobile solution is proposed to interpret scene texts of different orientations in natural scene images. Much research has been focused on automatic scene texts recognition. However, scene text orientations, image resolutions, various font sizes and more affect the accuracy of scene text recognition. Scene texts are commonly presented in many orientations, such as horizontal, arbitrary, curved and vertical. Research has recently focused on recognizing horizontal, arbitrary and curved scene texts but neglected the importance of recognizing vertically oriented scene texts (VOS). VOS can be differentiated into three types which are top-to-bottom (T2B) vertical scene texts, bottom-to-top (B2T) vertical scene texts and vertically stacked (VSt) vertical scene texts. Thus, 'Let Me Read for You' can detect and recognize types of VOS to access information from users' surroundings. A dataset, namely Vertically Oriented Scene Texts Dataset (VOD), consists of VOS images and is introduced for evaluation. The performance of 'Let Me Read for You' shows 86% accuracy in recognizing types of VOS, including T2B vertical texts, B2T vertical texts and VSt vertical texts, when evaluated with the VOD dataset.**

***Index Terms*—Mobile application, scene text recognition, text orientation, Vertical Text Interpreter.**

I. Introduction

AUTOMATIC scene text recognition has been a popular yet challenging research topic [1]. Scene texts appear in different font types, sizes, orientations, and colours, making scene text recognition more challenging. However, scene text recognition is important as it helps access information and benefits humans in many applications. For example, automatic scene text recognition provides navigation for people with visual impairment and helps understand natural semantic scenes, license plate recognition [2], etc.

Manuscript received October 06th, 2021; revised May 2nd, 2022. The Melbourne Sarawak Research Collaboration Scheme supported this research work.

Ong Yi Ling is a Senior Engineer of X-FAB Sarawak Sdn. Bhd. at Samajaya Free Industrial Zone, 93350 Kuching, Sarawak, Malaysia. (email: yileen_ong@hotmail.com)

Lau Bee Theng is a Director of the School of Research at Swinburne University of Technology Sarawak Campus, 93350 Kuching, Sarawak, Malaysia. (email: blau@swinburne.edu.my)

Almon Chai WeiYen is a Senior Lecturer at Swinburne University of Technology Sarawak Campus, 93350 Kuching, Sarawak, Malaysia. (email: achai@swinburne.edu.my)

Chris McCarthy is a Senior Lecturer and Industry Engagement Director in the School of Software and Electrical Engineering at Swinburne University of Technology Melbourne Campus, John St, Hawthorn VIC 3122, Australia. (email: cdmccarthy@swin.edu.au)

Therefore, scene text recognition models have been implemented in mobile applications, which are portable and low priced, so that users can understand scene texts from their surroundings and access information using the mobile applications on their smartphones.

In the past, the traditional Optical Character Recognition (OCR) method was introduced to detect texts on scanned documents [3]. Traditional OCR was also developed into mobile applications, such as 'OCR-Text Scanner' [4] and 'Image to Text' [5] applications. Both applications recognize handwritten and typed texts.

Comparing texts on documents and texts on natural scene images, texts in natural scenes are unpredictable with different text orientations, font types and complex backgrounds. Therefore, scene text recognition is way more challenging in natural scenes than constant texts on documents. Hence, the OCR method is not applicable for scene text recognition in natural images.

Scene text recognition is a famous research topic because there are many types of scene text orientation [6], such as the horizontal scene texts, arbitrarily oriented scene texts, curved scene texts and VOS [7]. In recent years, many researchers proposed different approaches for scene text recognition in natural images. However, most research focuses more on recognizing horizontal and oriented scene texts but not on VOS [8]. According to Liao, et al. [9], research is needed to look into VOS recognition modelling in natural scene images.

Recognizing VOS in natural scenes is as important as other text orientations because they carry information. Furthermore, there are three different VOS types: VSt vertical scene texts, B2T vertical scene texts, and T2B vertical scene texts. Fig. 1 shows examples of natural images with different kinds of VOS.

(a) (b) (c)

Fig. 1. VSt vertical scene texts (a), B2T vertical scene texts (b), and T2B vertical scene texts (c) in natural images.

Thus, a pervasive and ubiquitous mobile solution that detects and recognizes types of VOS in natural images named the 'Let Me Read for You' is designed and developed. VTI is used in the mobile solution to detect and recognize VSt vertical scene texts, B2T vertical scene texts and T2B vertical scene texts in real-time.

II. Related Works

In this developed world, computer technology has helped people with visual impairment perform their daily activities such as reading, communicating, searching and navigating independently. With the evolution of technology on mobile devices, many applications that run on mobile devices have become popular due to their low cost, portability and accessibility to information. Hence, several mobile applications are available for reading applications, object recognition and outdoor navigation. Therefore, 'Let Me Read for You' is an application proposed on mobile devices to recognize different orientations of scene texts in natural scenes.

Previously, the traditional OCR method was also implemented in mobile applications to detect texts on scanned documents, such as the 'OCR-Text Scanner' [4] and 'Image to Text' [5] application. Besides detecting texts on documents, mobile applications recognize texts in natural scenes that are needed, especially for people with visual impairment to access information in their surroundings. For example, Naver Papago [10] is a mobile application that recognizes scene texts in natural scenes. However, Naver Papago [10] can only recognize scene texts that are horizontally oriented. Therefore, 'Let Me Read for You' is a mobile application proposed in this research that recognizes scene texts of different orientations, including VOS in natural scene images.

In general, in the process of scene text recognition in natural images, both scene text detection and scene text recognition are important. The process of scene text recognition first detects the possible locations of text regions in natural images of scene text detection. After text regions are located, scene text recognition reads and converts the text regions into texts input for other purposes [11].

Scene text recognition begins with scene text detection. In the past years, the sliding window based method is one of the traditional methods, which implements a moving window to slide through the entire image to survey locations with a text [11]. Another conventional method, the connected component-based method, extracts text characters in the input image using extreme regions extraction and colour clustering [12]. For example, Huang, et al. [13] use stroke width transform (SWT) to differentiate text and non-text components. Kaushik and Verma [12] introduced a hybrid method that blends the benefit of two traditional methods mentioned for better performance of scene text detection.

As for scene text recognition converts detected text regions extracted by scene text detection into text strings that the computer can understand for other applications. The sequence-based method is one of the traditional methods used in scene text recognition. Using a sequence-based method, Shi, et al. [14] introduced Sequence Recognition Network (SRN), which implements an encoder-decoder framework to recognize scene texts. Bai, et al. [15] also introduced the edit probability (EP) method, which uses probability prediction to generate better accuracy of scene text recognition. Besides that, the word classification based method is another traditional method for scene text recognition. Kang, et al. [16] introduced a contextual text spotting network (CTSN) that implements learning words to improve its performance in recognition. Lee and Osindero [17] introduced a recursive recurrent neural network to learn character-level language models.

For better performances, deep learning-based methods were used for detecting and recognizing scene texts. For example, a cascaded convolutional text network (CCTN) is a deep learning-based method used to indicate the position of scene texts in natural images. He, et al. [18] managed to detect texts with arbitrary orientations and multi-languages using CCTN. Besides that, You Only Look Once (YOLO) algorithm was proposed by Qin and Manduchi [19] to identify text regions in an image using a fully convolutional network. Raisi, et al. [20] mentioned Connectionist Temporal Classification (CTC) is also suitable for extracting contextual information. Deshmukh [21] proved that CTC performs faster decoding than a Hidden Markov Model (HMM). Pelee-Text ++ is another lightweight neural network suggested by Pinto, et al. [22]. This network is implemented for any oriented scene text detection with different languages but not for scene text recognition.

Most research has recently focused on combining scene text detection and scene text recognition into a single model. TextBoxes++ is a successful model that works on recognizing both horizontal and arbitrary oriented scene texts in natural images using deep learning was introduced by Liao, et al. [9]. End-to-end Text Spotter is another model which detects and recognizes scene texts in a distinct neural network. This was proposed by He, et al. [23], which uses a single-shot text spotter to detect scene text regions while encoding and decoding series to recognize characters in the detected text regions.

Based on the findings, most researchers have been looking into recognizing horizontally oriented scene texts, arbitrarily oriented scene texts and curved scene texts. However, there is a lack of research for identifying VOS in scene images. Choi, et al. [24] proposed a model which recognizes horizontally oriented scene texts and VSt vertical scene texts. However, it fails to recognize B2T and T2B vertical scene texts in natural images. Besides that, Capture2Text [25] is a model which requires a manual selection and indication of the text region containing the vertically oriented Japanese characters, while the model only recognizes them. Capture2Text [25] also recognizes VSt vertical texts with 26 English alphabets, but it does not carry out detection.

According to Blanco Medina, et al. [26] and NguyenVan, et al. [27], VOS detection and recognition research lacks. Therefore, a mobile application named 'Let Me Read for You' is introduced in this research by implementing a VOS recognition model to overcome the gap of research in recognizing scene text. In addition, a VOS dataset is proposed in this research due to the lack of a dataset on VOS in natural images for evaluation purposes.

III. Design and Development of 'Let Me Read for You'

A. Design

'Let Me Read for You' is a mobile application that interprets scene texts in the natural scene images that the users capture. Therefore, users must first download this mobile application to their mobile phone. Once this application is ready on their mobile phone, users will get a welcoming message, as shown in Fig. 2.

Fig. 2. Welcoming message in 'Let Me Read for You'.

Next, the application will enter camera mode, which requires users to capture an image of their surroundings with scene texts, as demonstrated in Fig. 3.

Fig. 3. Camera mode in the application for image capturing.

Once the image with scene text is captured, users are asked to tap on the button 'DETECT' to proceed with scene text detection. Fig. 4 shows the instruction for detecting the scene texts in the application. Lastly, the results for the detected text in the images are shown in Fig. 6.

Fig. 5. Tap on the 'DETECT' button for detecting texts.

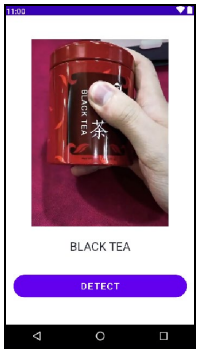


Fig. 6. Detected scene text results will be shown to the users.

B. Recognition Model

VTI is a model developed to detect and recognize three types of VOS, which are T2B vertical scene texts, B2T vertical scene texts and VSt vertical scene texts in 'Let Me Read for You'. VTI is the model proposed, which calls Vertical Text Spotter (VTS) first to detect VOS locations in a natural scene image. After the text regions are located, Vertical Text Reader (VTR) is introduced to read and recognize the VOS text regions. Fig. 7 shows the overall structure of VTI used in recognizing VOS.

VTI begins with a shared convolutions layer. Instead of using separate convolutional neural networks, a shared convolutional neural network is implemented for scene text detection and recognition. The shared convolutional neural network implements bottom-up and top-down pathways. First, the bottom-up pathway extracts the usual convolutional neural network feature maps. Then the top-down pathway provides a higher level of semantic feature map for better performance. Bottom-up and top-down pathways are combined by combining the previous layer and last layer feature map to obtain a feature map with higher resolution. The shared convolutional neural network is designed with ResNet50 and deconvolution layers. Using a shared convolutional network with bottom-up and top-down pathways, an up-sampling feature map is extracted for VTS and VTR.

VTS detects scene text of different orientations, including VOS. After feature maps are obtained from the shared convolutional layers, they are passed into VTS, which begins with Text Region Proposal Network (TRPN), as shown in Fig. 8. After that, the region proposed by TRPN undergoes Inclined Non-Maximum Suppression (INMS). Lastly, the rotation ROI pooling layer rotates the detected text regions into an axis-aligned text region.

In VTS, Text Region Proposal Network (TRPN) determines the possible location of text regions of different orientations, such as horizontal scene texts, arbitrarily oriented scene texts and VOS. TRPN is selected because it proposes text regions using rotated anchor boxes and can detect text regions with different orientations. TRPN uses rotated anchor boxes with different scales (8, 16 and 32), different aspect ratios (1:2, 1:5 and 1:8) and different angles of rotation ($-\pi/6$, 0, $\pi/6$, $\pi/3$, $\pi/2$ and $2\pi/3$) to propose possible text regions in the input images. A rotated anchor box is represented by five tuples, which are (x, y, h, w, θ), where (x, y) show the center of the anchor box, w and h represent the weight and height of the anchor box. θ showing the angle of rotation.

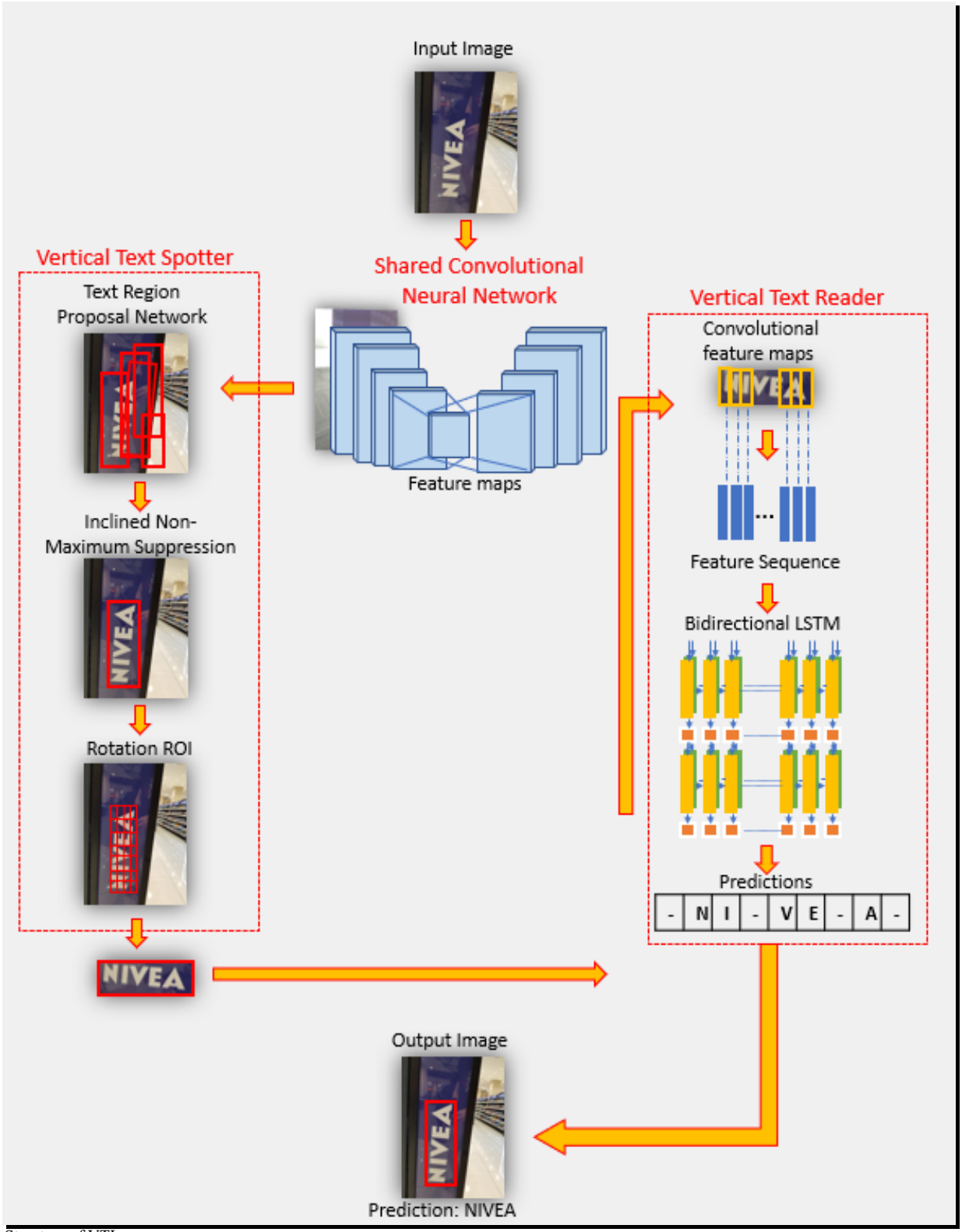


Fig. 7. Structure of VTI

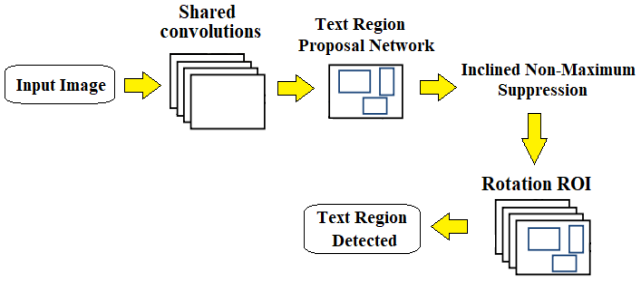


Fig. 8. Process flow of VTS.

Next, the possible text regions proposed by TRPN were passed on to Inclined Non-Maximum Suppression (INMS). In contrast to the traditional NMS, INMS considers arbitrarily oriented scene text region proposals. INMS is used when many overlapping bounding boxes are obtained for the detected text. Intersection over Union (IOU) is calculated to eliminate negatively detected bounding boxes [28]. IOU is calculated as the ground truth bounding box's overlapped area, and the area detected bounding box. This is divided by their union area in the calculation, as shown in Eq. (1).

Lastly, a detected rotated anchor box is considered a positively detected anchor box when the intersection angle is less than $\pi/12$, and the value of IOU is more than 0.7.

$$\frac{Area\ of\ Overlap}{Area\ of\ Union} = IOU \quad (1)$$

Finally, the correctly detected anchor boxes are fed into the Rotation ROI pooling layer. In rotation ROI pooling, the max-pooling algorithm is implemented on the sub-regions of the text region proposals; they can be rotated into axis-aligned text region proposals. Therefore, all detected text regions, such as horizontal scene text regions, arbitrarily oriented scene text regions and VOS text regions, are rotated into axis-aligned text regions as the output of VTS.

After text regions are detected, VTR is the next major process in VTI. Firstly, the extracted feature maps from shared convolutions and feature maps from VTS are both sent to VTR. Fig. 9 shows the overall process of VTR. VTR is implemented to recognize all three types of VOS – T2B vertical texts, B2T vertical texts and VSt vertical scene texts.

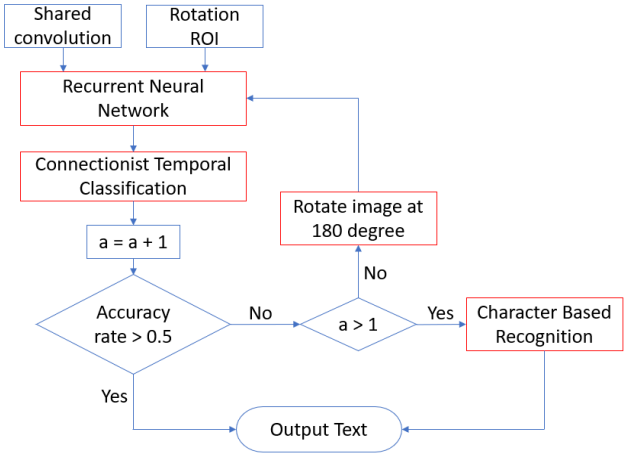


Fig. 9. Architecture of Vertical Text Reader.

The feature maps are first passed to a Recurrent Neural Network (RNN). The RNN model is used because it generates character sequences [29] and keeps temporal context information with the implementation of activation function, convolutional layers and max-pooling on the local region. In the RNN model, a feature sequence is created from the feature maps by moving the feature vectors from columns left to right. It suits VTR because it is fed with axis-aligned text regions from VTS. Fig. 10 shows the generated feature sequences correlate to the receptive field in the RNN model.

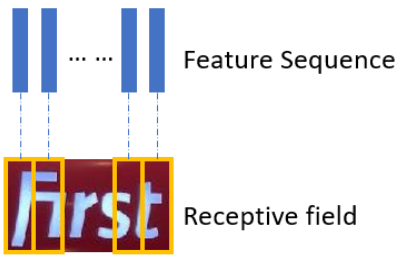


Fig. 10. An example of a feature sequence extracted from the receptive field.

Next, the feature sequence generated is fed into Bidirectional Long-Short Term Memory (LSTM). It is named bidirectional LSTM as it is made up of two LSTLSTMsichcaptures feature sequences and keeps them under hidden layers. Bidirectional LSTM considers both directions at each step in the calculation, which helps in better recognition performance.

After that, Connectionist Temporal Classification (CTC) is implemented to predict text labels in the detected text regions. CTC converts the feature sequences generated by RNN into label sequences. For example, the input sequence of CTC is shown by $y = y1, y2, y3, \ldots, yT$, where T shows the length of its sequence. An output label set, $L' = L \cup \{\text{blank}\}$ is formed, where L shows the true tags, while blank symbols show null prediction. Taking the example in Fig. 10, label sequences obtained show "- - f f – i – r s s – t t –". After that, it is converted into "first" using the sequence-to-sequence mapping function by removing the repeating labels and blank symbols. Eq. (2) shows the probability of the final output sequence obtained.

$$p(L|y) = \sum_{\pi:B(\pi)=1} p(\pi|y) \quad (2)$$

As shown in Fig. 9, the parameter 'a calculates the number of times the text regions undergo RNN and CTC. However, if the recognition rate is unsatisfactory, the detected text region is rotated 180° and undergoes RNN and CTC again. Character-based recognition will be further carried out to recognize VSt vertical scene texts.

The three types of VOS – VSt vertical texts, B2T vertical texts and T2B vertical texts - show different orientations of characters after rotation ROI. Fig. 11 also shows the character orientations of other VOS after they are rotated into axis-aligned text regions.

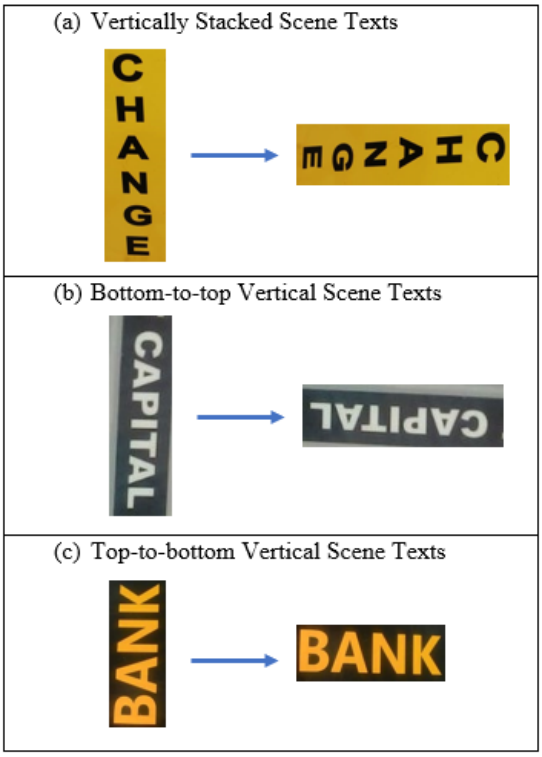


Fig. 11. Examples of VSt, B2T and T2B vertical scene texts rotated into axis-aligned text regions.

VTR works directly for B2T vertical texts as the characters in the text are horizontally oriented, it is not directly applicable for T2B vertical texts and VSt vertical scene texts. Therefore, when the recognition rate is less than 0.5 accuracy, axis-aligned text regions will be rotated 180° again. It helps recognize T2B vertical texts but not VSt vertical scene texts. Therefore, VSt vertical scene texts will undergo character-based recognition for further recognition. Fig. 12 shows the architecture of character-based recognition.

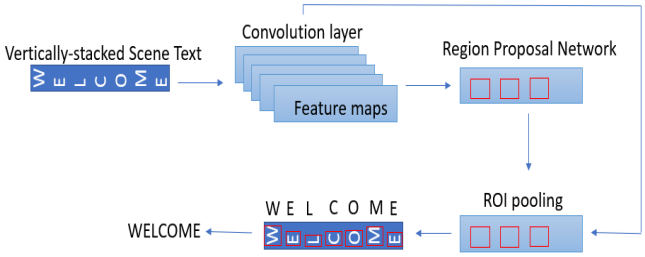


Fig. 12. Architecture of character-based recognition for VSt vertical scene texts.

Character-based recognition is required to recognize VSt vertical scene texts. Character-based recognition follows the idea proposed by Ren, et al. [30], who implemented a Faster RCNN model for object detection. The axis-aligned text region for VSt vertical scene texts goes through a convolution layer to extract feature maps. The feature maps are passed into the region proposal network to present text regions detected using anchor boxes. Anchor boxes with scales of 4, 8 and 16 and aspect ratios of 1:1, 1:2 and 2:1 are used, while ROI pooling removes negatively detected characters. Character-based recognition recognizes oriented alphabets and oriented numbers, as shown in Fig. 13.

Fig. 13. Character orientation in VSt vertical scene texts after rotation.

C. *Datasets for Training and Evaluation*

For evaluation purposes, International Conference on Document Analysis and Recognition (ICDAR) 2013 dataset [31] and ICDAR 2015 dataset [32] are selected as the benchmarking datasets. ICDAR 2013 and ICDAR 2015 datasets are benchmarking datasets used to evaluate scene text detection performance.

ICDAR 2013 dataset is a popular dataset with a focused scene text dataset, where images in the dataset focus on horizontal scene texts, as shown in Fig. 14 (a). There are 229 training images and 233 testing images [31]. Besides that, the ICDAR 2015 dataset is another benchmarking dataset selected. It is known as the Incidental Scene Text dataset, which consists of mostly arbitrary oriented texts in scene images, as shown in Fig. 14 (b). In ICDAR 2015 dataset, the training set consists of 1,000 images, while the testing set consists of 500 images.

(a) (b)

Fig. 14. Examples of natural scene images in ICDAR 2013 (a) and ICDAR 2015 (b) datasets.

Besides that, the IIIT 5K-Word dataset is used as the benchmarking dataset for evaluating the accuracy of scene text recognition. IIIT 5K-Word dataset contains 5,000 images with both digital and natural scene images. According to Mishra, et al. [33], the largest dataset to date, the IIIT 5K-Word dataset, is useful for scene text recognition evaluation. Sample images for IIIT 5K-Word datasets are shown in Fig. 15.

Fig. 15. Sample images for IIIT 5K-Word dataset.

However, there is still a lack of a dataset for VOS. Therefore, a Vertically Oriented Scene Text Dataset (VOD) is proposed in this research. VOD dataset is developed to evaluate the performance of VTI in detecting and recognizing different types of VOS. In the VOD dataset. There are 1,000 images with types of VOS collected, where the training set has 800 pictures and the testing set has 200 images. Examples of natural scene images in the VOD dataset with different types of VOS are shown in Fig. 16.

Fig. 16. Natural scene images in the VOD dataset with different types of VOS.

IV. Results and Discussion

Therefore, VTI is a VOS detection and recognition model proposed in this research. In developing VTI, several phases of experimental testing have been carried out. There was a total of three phases of development in this research project. The first phase was the evaluation of VTS in detecting scene text regions. The second phase was the evaluation of VTR in recognizing VOS text regions. Lastly, the third phase was the evaluation of VTI, where the overall end-to-end vertical scene text recognition model is evaluated. Finally, VTI is implemented into a mobile application, 'Let Me Read for You', so that scene text recognition can be used anywhere due to its portability by using a mobile phone, which helps users access information everywhere they go.

In the first phase, the performance of VTS is evaluated to determine its accuracy in detecting the presence of VOS in natural scene images. With the implementation of rotation anchor boxes in VTS, VTS detects VOS and is also capable of detecting horizontally oriented scene texts and arbitrarily oriented scene texts.

Therefore, several benchmarking datasets, ICDAR 2013, ICDAR 2015 and VOD datasets, are used to evaluate the performance of VTS in detecting different scene text orientations. Since ICDAR 2013 dataset is called as Focus Scene Text Dataset with horizontally oriented scene texts [31], ICDAR 2013 dataset is used to evaluate the performance of VTS in detecting horizontally oriented scene texts. At the same time, ICDAR 2015 dataset is used to assess the performance of VTS in detecting arbitrarily oriented scene texts, while the VOD dataset is used to evaluate the performance of VTS in detecting VOS.

Fig. 17. Sample results of scene text detection using VTS with the ICDAR 2013 dataset (a), ICDAR 2015 dataset (b) and VOD dataset (c), where the arrow shows the boxes of detected scene texts region.

To evaluate the performance of VTS, the evaluation metrics used for evaluation are precision, recall and f-measure metrics. Precision calculates the percentage of correctly detected text regions compared to the ground truth text regions, as shown in Eq. (3). The recall metric determines the sensitivity of VTS in detecting text regions, as shown in Eq. (4). Eq. (5) shows that the f-measure takes the harmonic average of precision and recall metrics.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (3)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (4)$$

$$F - measure = \frac{2\ x\ Precision\ x\ Recall}{Precision + Recall} \quad (5)$$

The evaluation results of VTS with ICDAR 2013, ICDAR 2015 and VOD datasets are tabulated and compared with other approaches, as shown in Table 1. Sample results of VTS in detecting scene text regions in natural scene images of ICDAR 2013, ICDAR 2015 and VOD datasets are shown in Fig. 17.

The results obtained in Table 1 show that VTS shows equivalent performance to the other state-of-the-art methods in detecting different orientations of scene texts, including horizontally oriented scene texts, arbitrarily oriented scene texts and VOS.

TABLE I
PERFORMANCE OF VTS USING ICDAR 2013, ICDAR 2015 AND VOD DATASETS IN DETECTING SCENE TEXT REGIONS IN NATURAL SCENE IMAGES.

Methods	ICDAR 2013			ICDAR 2015			VOD		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
VTS	0.91	0.89	0.89	0.84	0.82	0.83	0.87	0.76	0.81
Wang, et al. [34]	0.84	0.83	0.81	0.84	0.82	0.82	-	-	-
Long, et al. [35]	0.83	0.74	0.78	0.85	0.80	0.82	-	-	-
He, et al. [36]	0.92	0.81	0.86	0.82	0.80	0.81	-	-	-
Shi, et al. [37]	0.87	0.83	0.85	0.73	0.77	0.75	-	-	-
Tian, et al. [38]	0.93	0.83	0.88	0.74	0.52	0.61	-	-	-

In the second phase of this research, the performance of VTR in recognizing VOS is evaluated. VTR was assessed on its performance in identifying scene texts on cropped word images. At this stage, VTR was evaluated on its accuracy in recognizing texts detected in text regions. For text recognition, the recognition rate is used to assess the performance of VTR by calculating the percentage of the number of words correctly recognized to the total number of words, as shown in Eq. (6).

$$Recognition\ Rate = \frac{No.\ of\ words\ correctly\ recog.}{Total\ no.of\ words}\ x\ 100\% \quad (6)$$

VTR was not only capable of recognizing VSt vertical scene texts, T2B vertical scene texts and B2T vertical scene texts but also recognizing horizontally oriented scene texts. VTR was evaluated with the IIIT 5K-Word and ICDAR 2013 datasets suitable for scene text recognition evaluation, and the VOD dataset consists of three different types of VOS cropped word images. The results on the accuracy of scene text recognition are recorded and compared to other approaches in Table 2.

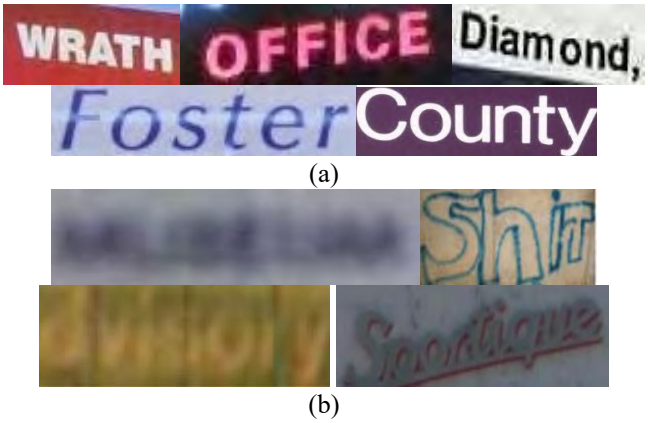
TABLE II
RECOGNITION RATE OF VTR ON IIIT 5-K WORD, ICDAR 2013 AND VOD DATASETS IN RECOGNIZING SCENE TEXTS.

Methods	IIIT 5K-Word	ICDAR 2013	VOD
VTR	0.81	0.90	0.85
Wang, et al. [39]	0.90	-	-
Liao, et al. [40]	0.92	0.91	-
Li, et al. [41]	0.92	0.91	-
Cheng, et al. [42]	0.87	-	-
Liu and Lane [43]	0.78	0.91	-

As shown in Table 2, the VTR performs equivalently to other approaches. Fig. 18 (a) shows examples of successful recognition from IIIT 5K-Word, ICDAR 2013 and VOD dataset, while Fig. 18 (b) shows unsuccessful recognition cases. The unsuccessful cases are mainly blurry images and cursive text font types.

The performance of VTI was evaluated in the third phase of this research. The performance of VTI, combining both VTS and VTR, is evaluated. Evaluation will match the detection result with ground truth instances and later compare the recognized text content.

In this phase, the VOD dataset evaluates the performance of VTI in recognizing VOS. Besides, benchmarking datasets of ICDAR 2013 and ICDAR 2015 are also used for evaluation purposes. ICDAR 2013 dataset evaluates the result of VTI in identifying horizontally oriented scene texts, while the ICDAR 2015 dataset is used to assess recognizing arbitrarily oriented scene texts. Table 3 shows the evaluation results of VTI in recognizing horizontally oriented scene texts, arbitrary oriented scene texts and VOS using ICDAR 2013, ICDAR 2015 datasets and VOD datasets.



(a)

(b)

Fig. 18. Examples of successful (a) and unsuccessful (b) recognition cases from IIIT 5K-Word, ICDAR 2013 and VOD datasets.

Based on Table 3, the result of VTI in recognizing VOS is evaluated with the VOD dataset in the third phase. VTI shows an accuracy of 86% rate of recognition on identifying VOS. Besides that, VTI achieved a 62% recognition rate on the ICDAR 2015 dataset and 84% recognition on the ICDAR 2013 dataset.

TABLE III
RESULT OF VTI ON VOD, ICDAR 2015 AND ICDAR 2013 DATASETS, WHERE 'P' SHOWS PRECISION, 'R' SHOWS RECALL, 'F' SHOWS F-MEASURE AND 'RECOG.' SHOWS THE RATE OF RECOGNITION.

Methods	Dataset											
	VOD				ICDAR 2015				ICDAR 2013			
	P	R	F	Recog.	P	R	F	Recog.	P	R	F	Recog.
VTI	0.87	0.76	0.81	0.85	0.85	0.83	0.84	0.62	0.91	0.88	0.89	0.84
Wang, et al. [39]				-	0.88	0.87	0.88	0.64	0.89	0.85	0.87	0.84
Qiao, et al. [44]				-	0.91	0.81	0.86	0.63	0.92	0.88	0.90	0.84
Lyu, et al. [45]				-	0.86	0.81	0.86	0.62	0.95	0.88	0.91	0.86
Liao, et al. [9]				-	0.87	0.77	0.82	0.52	0.74	0.86	0.80	0.85
He, et al. [23]				-	0.83	0.84	0.83	0.63	0.91	0.88	0.90	0.86

Comparing the result of VTI with Wang, et al. [39], Qiao, et al. [44], Lyu, et al. [45], Liao, et al. [9] and He, et al. [23], as shown in Table 1. These methods are selected for comparison as they perform scene text detection and recognition on horizontally oriented scene texts and arbitrary oriented scene texts. The comparison proves VTI performs equivalently on recognizing horizontal scene texts and arbitrary oriented scene texts. Furthermore, VTI simultaneously recognizes B2T vertical scene texts, T2B vertical scene texts, and VSt vertical scene texts. Samples of scene text recognition that VTI successfully carried out are shown in Fig. 19.

Fig. 19. Recognition of VOS, arbitrary oriented scene texts and horizontally oriented scene texts by VTI.

Finally, VTI is implemented into a mobile application, 'Let Me Read for You'. With this mobile application, users can access different orientations of natural scene texts in their surroundings. Fig. 20 below shows sample results obtained using 'Let Me Read for You'.

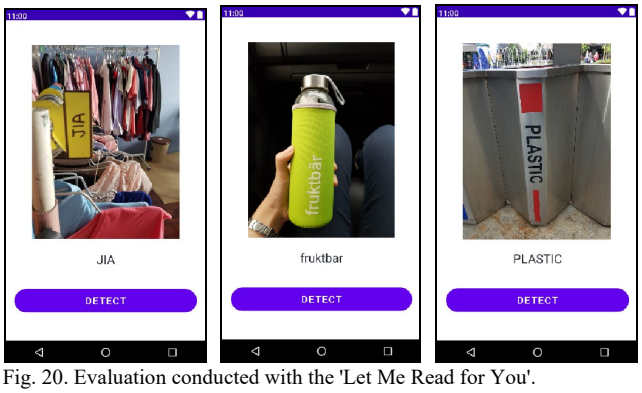


Fig. 20. Evaluation conducted with the 'Let Me Read for You'.

V. Conclusions and Future Works

There are different text orientations in natural images, such as horizontal, arbitrary, curved and VOS. In recent years, most research has focused on recognizing horizontal, arbitrary and curved scene texts. However, limited study on recognizing VOS has been carried out. Therefore, an end-to-end VOS recognition model, namely VTI, is proposed in this research. VTI is capable of recognizing three types of VOS, which are B2T vertical scene texts, T2B vertical scene texts and VSt vertical scene texts. VTI merges VTS and VTR. VTS performs scene text detection, while VTR recognizes VOS. Lastly, character-based recognition is implemented on VTR in recognizing VSt vertical scene texts. Besides that, a dataset that consists of types of VOS, namely the VOD dataset, is presented in this research. VTI is evaluated for its detection and recognition performance using VOD, IIIT 5K-Word, ICDAR 2015 and ICDAR 2013 datasets. Evaluation proves that VTI can simultaneously recognize horizontally oriented scene texts, arbitrarily oriented scene texts, and VOS. For better accessibility to information, 'Let Me Read for You' is developed with the VTI model's implementation in interpreting different orientations of scene texts in natural scenes.

For future works, the VTI can be adapted into self-driving cars technology. Besides that, the 'Let Me Read for You' mobile application can be further improved into an orientation and mobility assistance application for the visually impaired.

References

[1] P. X. Nguyen, K. Wang, and S. Belongie, "Video text detection and recognition: Dataset and benchmark," in IEEE Winter Conference on Applications of Computer Vision 2014, 24-26 March 2014, USA, pp. 776-783, 2014, doi: 10.1109/WACV.2014.6836024.

[2] G. Kosala, A. Harjoko, and S. Hartati, "Robust License Plate Detection in Complex Scene using MSER-Dominant Vertical Sobel," *IAENG International Journal of Computer Science,* vol. 47, no. 2, pp. 214-222, 2020. Available at: http://www.iaeng.org/IJCS/issues_v47/issue_2/IJCS_47_2_10.pdf.

[3] S. F. Rashid, "Optical Character Recognition-A Combined ANN/HMM Approach", thesis, Technical University of Kaiserslautern, Germany, 2014. Available at: https://kluedo.ub.uni-kl.de/files/3939/rashid-thesis_11July_2014.pdf.

[4] "OCR-Text Scanner," ed. Google Play. Available at: https://play.google.com/store/apps/details?id=com.offline.ocr.english.image.to.text

[5] "Image to Text," ed: Google Play. Available at: https://play.google.com/store/apps/details?id=ngapha m.com.vnocr&hl=en

[6] Y. Chen and J. Yang, "Research on Scene Text Recognition Algorithm Basedon Improved CRNN," in Proceedings of the 2020 4th International Conference on Digital Signal Processing, pp. 107-111, 2020, doi: https://doi.org/10.1145/3408127.3408166.

[7] O. Y. Ling, L. B. Theng, A. Chai, and C. McCarthy, "A Model for Automatic Recognition of Vertical Texts in Natural Scene Images," in *2018 8th IEEE International Conference on Control System, Computing and Engineering (ICCSCE)*, 23-25 Nov, Penang, Malaysia, pp. 170-175, 2018, doi: 10.1109/ICCSCE.2018.8685019, 2018.

[8] B. T. L. Yi Ling Ong, Almon Chai. Chris McCarthy, "A Deep Learning Framework for Recognizing Vertical Texts in Natural Scene," presented at the 2nd International Conference on Computer and Drone Applications 2019, Sarawak, Malaysia, 19-21 December, pp. 144341-144351, 2019, doi: 10.1109/IConDA47345.2019.9034915.

[9] M. Liao, B. Shi, and X. Bai, "Textboxes++: A single-shot oriented scene text detector," IEEE transactions on image processing, vol. 27, no. 8, pp. 3676-3690, 2018, doi: 10.1109/TIP.2018.2825107.

[10] "Naver Papago," ed. Google Play. Available at: https://play.google.com/store/apps/details?id=com.naver.labs.translator&hl=en&gl=US

[11] H. Lin, P. Yang, and F. Zhang, "Review of Scene Text Detection and Recognition," Archives of Computational Methods in Engineering, 11 January 2020, pp. 433-454, 2020, doi: 10.1007/s11831-019-09315-1.

[12] D. Kaushik and V. S. Verma, "Review on Text Recognition in Natural Scene Images," in Innovations in Computational Intelligence: Springer, pp. 29-43, 2018, doi: 10.1007/978-981-10-4555-4_3.
[13] W. Huang, Z. Lin, J. Yang, and J. Wang, "Text localization in natural images using stroke feature transform and text covariance descriptors," in Proceedings of the IEEE International Conference on Computer Vision, pp. 1241-1248, 2013. Available at: https://openaccess.thecvf.com/content_iccv_2013/papers/Huang_Text_Localization_in_2013_ICCV_paper.pdf.
[14] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4168-4176, 2016. Available at: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Shi_Robust_Scene_Text_CVPR_2016_paper.pdf.
[15] F. Bai, Z. Cheng, Y. Niu, S. Pu, and S. Zhou, "Edit probability for scene text recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1508-1516, 2018, doi: 10.48550/arXiv.1805.03384.
[16] C. Kang, G. Kim, and S. I. Yoo, "Detection and recognition of text embedded in online images via neural context models," in Thirty-First AAAI Conference on Artificial Intelligence, pp. 4103-4110, 2017. Available at: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14154/14287.
[17] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for ocr in the wild," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2231-2239, 2016.
[18] T. He, W. Huang, Y. Qiao, and J. Yao, "Accurate text localization in natural image with cascaded convolutional text network," arXiv preprint arXiv:1603.09423, 2016.
[19] S. Qin and R. Manduchi, "Cascaded segmentation-detection networks for word-level text spotting," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1: IEEE, pp. 1275-1282, 2017, doi: 10.1109/ICDAR.2017.210.
[20] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, and J. Zelek, "Text Detection and Recognition in the Wild: A Review," arXiv preprint arXiv:2006.04305, 2020.
[21] A. M. Deshmukh, "Comparison of Hidden Markov Model and Recurrent Neural Network in Automatic Speech Recognition," European Journal of Engineering and Technology Research, vol. 5, no. 8, pp. 958-965, 2020, doi: 10.24018/ejeng.2020.5.8.2077.
[22] A. Pinto, H. Pedrini, and R. da Silva Torres, "Pelee-Text++: A Tiny Neural Network for Scene Text Detection," IEEE Access, vol. 8, pp. 223172-223188, 2020, doi: 10.1109/ACCESS.2020.3043813.
[23] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, and C. Sun, "An end-to-end textspotter with explicit alignment and attention," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5020-5029, 2018. Available at: https://openaccess.thecvf.com/content_cvpr_2018/papers/He_An_End-to-End_TextSpotter_CVPR_2018_paper.pdf.
[24] C. Choi, Y. Yoon, J. Lee, and J. Kim, "Simultaneous Recognition of Horizontal and Vertical Text in Natural Images," in Asian Conference on Computer Vision, pp. 202-212, 2018, doi: 10.1007/978-3-030-21074-8_16.
[25] Capture2Text, "Capture2Text", 2018. Available at: <http://capture2text.sourceforge.net/> [Accessed: 15 August 2018].
[26] P. Blanco Medina, E. Fidalgo, E. Alegre, M. W. Al-Nabki, and D. Chaves, "Enhancing text recognition on Tor Darknet images," in XL Jornadas de Automática, Universidade da Coruña, Servizo de Publicacións, pp. 828-835, 2019, doi: 10.17979/spudc.9788497497169.
[27] D. NguyenVan, S. Lu, S. Tian, N. Ouarti, and M. Mokhtari, "A pooling based scene text proposal technique for scene text reading in the wild," Pattern Recognition, vol. 87, pp. 118-129, 2019, doi: 10.1016/j.patcog.2018.10.012.
[28] Y. Ling Ong, B. Theng Lau, A. Chai, and C. McCarthy, "Detecting of Vertically-Oriented Texts in Images Containing Natural Scenes," in MobiQuitous 2020-17th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services, pp. 444-450, 2020, doi: 10.1145/3448891.3450336.
[29] S. Fang, H. Xie, Z.-J. Zha, N. Sun, J. Tan, and Y. Zhang, "Attention and Language Ensemble for Scene Text Recognition with Convolutional Sequence Modeling," Proceedings of the 26th ACM international conference on Multimedia, Seoul, Republic of Korea, 2018, doi: 10.1145/3240508.3240571.
[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, pp. 91-99, 2015. Available: https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.
[31] D. Karatzas *et al.*, "ICDAR 2013 robust reading competition," in 2013 12th International Conference on Document Analysis and Recognition, pp. 1484-1493, 2013, doi: 10.1109/ICDAR.2013.221.
[32] D. Karatzas *et al.*, "ICDAR 2015 competition on robust reading," in *2015 13th International Conference* on Document Analysis and Recognition (ICDAR), pp. 1156-1160, 2015, doi: 10.1109/ICDAR.2015.7333942.
[33] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in BMVC-British Machine Vision Conference, 2012, doi: 10.5244/C.26.127**.**
[34] X. Wang, Y. Jiang, Z. Luo, C.-L. Liu, H. Choi, and S. Kim, "Arbitrary shape scene text detection with adaptive text region representation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6449-6458, 2019. Available at: https://openaccess.thecvf.com/content_CVPR_2019/papers/Wang_Arbitrary_Shape_Scene_Text_Detection_With_Adaptive_Text_Region_Representation_CVPR_2019_paper.pdf.
[35] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in Proceedings of the European conference on computer vision (ECCV), pp. 20-36, 2018, doi: 10.48550/arXiv.1807.01544.
[36] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in Proceedings of the IEEE International Conference on Computer Vision, pp. 745-753, 2017, doi: 10.48550/arXiv.1703.08289.
[37] B. Shi, X. Bai, and S. Belongie, "Detecting oriented text in natural images by linking segments," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2550-2558, 2017, doi: 10.48550/arXiv.1703.08289.
[38] Z. Tian, W. Huang, T. He, P. He, and Y. Qiao, "Detecting text in natural image with connectionist text proposal network," in European conference on computer vision, Springer, pp. 56-72, 2016, doi: 10.1007/978-3-319-46484-8_4.
[39] H. Wang *et al.*, "All you need is boundary: Toward arbitrary-shaped text spotting," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12160-12167, 2020, doi: 10.1609/aaai.v34i07.6896.
[40] M. Liao *et al.*, "Scene text recognition from two-dimensional perspective," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 8714-8721, 2019, 10.1609/aaai.v33i01.33018714.
[41] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 8610-8617, 2019, doi: 10.1609/aaai.v33i01.33018610.
[42] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu, and S. Zhou, "Aon: Towards arbitrarily-oriented text recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5571-5579, 2018, doi: 10.48550/arXiv.1711.04226.
[43] B. Liu and I. Lane, "An end-to-end trainable neural network model with belief tracking for task-oriented dialog," published at Interspeech 2017, 20 August 2017, doi: 10.21437/Interspeech.2017-1326.
[44] L. Qiao *et al.*, "Text perceptron: Towards end-to-end arbitrary-shaped text spotting," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 11899-11907, 2020, doi: 10.1609/aaai.v34i07.6864.
[45] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in Proceedings of the European Conference on Computer Vision (ECCV), pp. 67-83, 2018.

Ong Yi Ling received her Bachelor of Engineering (Hons) degree majoring in Electrical and Electronics from the Swinburne University of Technology, Sarawak Campus, Malaysia. She was a post-doc researcher and previously worked as a teaching assistant in Swinburne Sarawak while pursuing her PhD degree. Her research interests are image processing and text recognition system in assisting people with visual impairment.

Lau Bee Theng has actively contributed to her research areas with various edited books, peer-reviewed journals, conference proceedings, higher degrees in research, and funded research projects. Her research interests include artificial intelligence in activity recognition, natural scene text recognition, speech emotion detection, road accidents recognition, wafer surface defect detection, financial risks recognition, aesthetic preference of design objects, etc.

Almon Chai WeiYen discovered that energy conservation could be performed on the roller kilns' drying process, pursued further research, and commenced his doctoral study on computational modelling and simulations.

Almon has also managed a transport innovation centre for a few years before 2021, with projects ranging from IoT implementation in automotive component manufacturing to component manufacturing process simulation. Almon has research interests in computational simulations, modelling, IoT implementation, mechatronics and human factors.

Chris McCarthy is a Senior Lecturer and Industry Engagement Director in the School of Software and Electrical Engineering at the Swinburne University of Technology. His research is in computer vision algorithms for robotics, intelligent transport systems, and human-computer interaction (particularly for assistive technologies). Dr McCarthy received his PhD from the Australian National University in 2010, during which time he worked on a novel biologically-inspired computer vision for robot navigation. Dr McCarthy has been at the forefront of computer vision for assistive technologies. He has developed and tested novel computer vision and visualization algorithms to support enhanced functional outcomes in blind patients implanted with a retinal prosthesis.