Journal of Electrical and Computer Engineering Innovations
(JECEI)
Journal homepage: http://www.jecei.sru.ac.ir

Research paper

Text Detection and Recognition for Robot Localization

Z. Raisi [1,2,*], *J. Zelek* [2]

[1]*University of Waterloo, Waterloo, Canada and Chabahar Maritime University, Chabahar, Iran.*

[2]*Systems Design Engineering Department, University of Waterloo, Canada.*

Article Info

Article History:
Received 26 June 2023
Reviewed 13 August 2023
Revised 05 September 2023
Accepted 08 September 2023

Keywords:
Text detection
Text recognition
Robot localization
Deep learning
Visual place recognition

*Corresponding Author's Email Address: *zraisi@uwaterloo.ca*

Abstract

Background and Objectives: Signage is everywhere, and a robot should be able to take advantage of signs to help it localize (including Visual Place Recognition (VPR)) and map. Robust text detection & recognition in the wild is challenging due to pose, irregular text instances, illumination variations, viewpoint changes, and occlusion factors.

Method: This paper proposes an end-to-end scene text spotting model that simultaneously outputs the text string and bounding boxes. The proposed model leverages a pre-trained Vision Transformer (ViT) architecture combined with a multi-task transformer-based text detector more suitable for the VPR task. Our central contribution is introducing an end-to-end scene text spotting framework to adequately capture the irregular and occluded text regions in different challenging places. We first equip the ViT backbone using a masked autoencoder (MAE) to capture partially occluded characters to address the occlusion problem. Then, we use a multi-task prediction head for the proposed model to handle arbitrary shapes of text instances with polygon bounding boxes.

Results: The evaluation of the proposed architecture's performance for VPR involved conducting several experiments on the challenging Self-Collected Text Place (SCTP) benchmark dataset. The well-known evaluation metric, Precision-Recall, was employed to measure the performance of the proposed pipeline. The final model achieved the following performances, Recall = 0.93 and Precision = 0.8, upon testing on this benchmark.

Conclusion: The initial experimental results show that the proposed model outperforms the state-of-the-art (SOTA) methods in comparison to the SCTP dataset, which confirms the robustness of the proposed end-to-end scene text detection and recognition model.

Introduction

We live in a visual world; signage is everywhere. Whether it is a street sign, a billboard, a house or room number, or labels such as a license plate or a person's name, signage provides us useful information in terms of location and identity. There have been many classifiers developed that are able to identify street signs or license plates with highly constrained priors on the method that do not allow their extension to general text in the wild detection and recognition.

However, to take advantage of all the signage available, we need to be able to detect signage (i.e., text) anywhere (i.e., in the wild). OCR is a well-solved problem for text detection and recognition in highly constrained environments; however, detecting and recognizing text anywhere is a challenging problem.

Signage can help a robot localize or map an environment. Typically, for SLAM processes, direct (i.e., pixel) or indirect features are used. Signage can provide a coarse localization globally when the signage indicates an

address or location. Also, the letters and numbers in a sign and their perspective can be used to determine relative pose if it can be assumed the signage is on a planar surface or even just vertical with respect to the ground plane. Visual Place Recognition (VPR) [15], [22], [41], [44], [69] aims to aid a vision-guided system to localize with respect to a previously visited place. VPR has uses in loop closure detection for visual SLAM and localization in general. Challenges in VPR include appearance variation due to perceptual aliasing, illumination, viewpoint changes, pose, weather, and seasons, to name a few. Most techniques are focused on features (i.e., indirect) [29] and sets of these features (e.g., BOW Bag of Words) methods.

Text spotting in wild images is also called end-to-end scene text detection and recognition [40], [52]. Simultaneous text detection and recognition go hand in hand. In scene text detection, the goal is to localize words in the image, and for scene text recognition, the aim is to convert the patch of cropped word images into a sequence of characters. Like scene text detection and recognition tasks, scene text spotting also encounters different challenging problems, including irregular text, illumination variations, low-resolution text, occlusion, *etc* [50].

Previous methods in scene text detection and recognition have utilized a convolutional neural network (CNN) as a feature extractor [19], [37], [56], [57] and Recurrent Neural Networks (RNN) [4], [21], [59] for capturing sequential dependency. Despite achieving promising performances on various challenging benchmark datasets [9], [16], [23]-[26], [35]-[42], [46], [48], [58], [60], [65]-[67], it has been shown that there are two main challenges for detecting or recognizing text in the wild images that have been studied in the past years. (1) Irregular text refers to text with arbitrary shapes that usually have severe orientation and curvature, and (2) occlusion, which makes poor performance on the existing methods [4], [5], [61] due to their reliance on the visibility of the target characters in the given images. Furthermore, CNNs have two significant drawbacks: (1) they have problems in capturing long-range dependencies (e.g., arbitrary relations between pixels in spatial domains) due to their fixed-size window operation [70], (2) they suffer from dynamical adoption to the changes to the inputs because the convolution filter weights are tuned to a specific training distribution [27].

Recent end-to-end scene text spotting methods [28], [54], [55], [58] utilized transformers [64] in their architecture and achieved superior performance in many benchmarks [9], [67]. Transformers [64] and their variations [7], [10], [70] are a new deep-learning architecture that mitigates the issues mentioned above for CNNs. Unlike Recurrent Neural Networks (RNNs), transformers are models that learn how to encode and decode data by looking not only backward but also forward to extract relevant information from a whole sequence, allowing conducting complex tasks such as machine translation [64], speech recognition [8], and recently, computer vision [7], [12], [27]. The attention mechanism allows the transformers to reason more effectively and focus on the relevant parts of the input data (e.g., a word in a sentence for machine translation and a character of a word in a text image for detection and recognition) as needed.

Visual place recognition (VPR) [22] aims to recognize previously visited places using visual information with resilience to perceptual aliasing, illumination, and viewpoint changes. Most of the techniques in VPR used keypoint features such as corners, edges, or blobs to represent and match distinctive points between the images. However, many keypoint features are needed to extract from images to establish a robust and repeatable representation of places and facilitate reliable localization and mapping in various applications like autonomous navigation, augmented reality, and robot localization. This process can be expensive in terms of computation and matching. On the other hand, as shown in Fig. 1, text features are semantic indexes and fewer in number compared to point features. Text instances that appear in the wild images, such as street signs, billboards, and shop signage, usually carry extensive discriminative information. VPR task can take advantage of these scene texts with high-level information for previously visited place recognition.

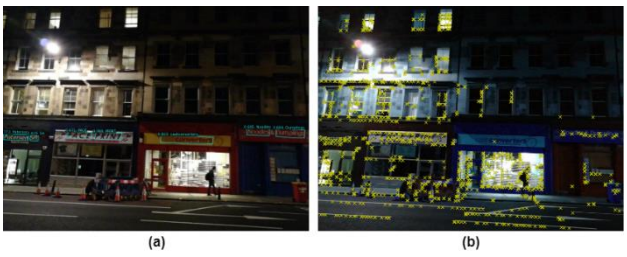

Fig. 1: Comparing the (a) Text features used in the proposed E2E text detection model and (b) Key point features used in different VPR techniques [11], [45]. Text features are shown with 'cyan' color boxes, and the 'x' marks with 'yellow' color denote the keypoint features.

This paper leverages a pre-trained end-to-end transformer-based text spotting framework for the VPR task. Unlike [22], which used two separate modules of detection and recognition for extracting the text regions, the final model can directly read the text instances from the given frame in an end-to-end manner. Furthermore, by equipping a masked autoencoder (MAE) [18] as a backbone, the proposed model is more robust in

capturing occluded text instance regions, which makes it more suitable for visual place recognition and other applications, including assistive technology for visually impaired people, autonomous vehicles, automated translation, and language processing in the wild images, information extraction from videos, and mobile applications for OCR and text, to name a few [36], [50].

The main contributions of this paper are as follows: (1) it utilizes an end-to-end transformer-based scene text spotting pipeline for the VPR application for the first time. The main difference between this method and other approaches is that it can directly output semantic text features (word instances and their bounding boxes), much less than the keypoint features used in most VPR techniques. (2) The proposed model utilizes a modified version of ViT by leveraging masked as input and adding a multi-scale adapter at the output to extract suitable features later for detection and recognition. (3) At the final stage, after utilizing a transformer-based detection architecture, this work uses a prediction head capable of simultaneously detecting the characters in the input image with their predicted classes and the bounding boxes of the word instances in the image. (4) By joining the middle point of the detected characters and their classes, the proposed model can handle arbitrary shapes text and output polygon bounding boxes with their word instances simultaneously, which makes it suitable not only for VPR but for other several text detection and recognition based in the wild applications. (5) This work also provides several quantitative and qualitative comparisons of the proposed technique with state-of-the-art (SOTA) in both VPR and scent text techniques.

Related Work

Scene text spotting aims to detect and recognize text instances from a given image end-to-end [14], [31], [33], [38], [39], [43], [47], [57]. Like different computer vision tasks, deep learning techniques using CNN/RNN-based methods and transformer-based methods are dominant frameworks in scene text spotting.

Early methods [31], [38] in scene text spotting have mainly utilized a deep-learning convolutional neural network (CNN) as a feature extractor [20] and Recurrent Neural Networks (RNN) [4], [21], [59] to read horizontal scene text. For example, Li *et al.* [31] combined the detection and recognition framework to present the first text-spotting method by using a shared CNN backbone encoder, followed by RoIPooling [57] as detection. Then, the resulting features are fed into the RNN recognition module to output the final word instances for a given input image. FOTS [38] utilized an anchor-free CNN-based object detection framework that improved both the training and inference time. It also uses RoIRotate for reading rotated text instances.

Since the text in the wild images appears in arbitrary shapes, including multi-oriented and curved, several methods [14], [33], [39], [47] targeted reading these types of text instances. These methods usually used a CNN-based segmentation network with multiple post-processing stages to output polygon box coordinates for the final irregular texts. For instance, in [47], a RoiMask is used to connect both the detection and recognition modules for capturing arbitrarily shaped text. Liu [39] leveraged a Bezier curve representation for the detection part, followed by a Bezier Align module to rectify the curved text instances into a regular text before feeding it to the attention-based recognition part. Some methods [6], [55] targeted spotting individual characters and merging them to output the final arbitrary shape text instance.

The Transformer framework, introduced by Vaswani et al. [64] for natural language processing tasks, has become a foundational architecture in various domains, including computer vision. In natural language processing, the Transformer architecture was initially designed to handle sequential data, such as sentences. It introduces a novel self-attention mechanism that allows the model to weigh the importance of different input elements when generating outputs. This attention mechanism enables parallel processing of sequences and mitigates the limitations of recurrent neural networks (RNNs) and convolutional neural networks (CNNs) in handling long-range dependencies.

Inspired by the success of Transformers in natural language processing, Dosovitskiy et al. [12] extended the Transformer framework to computer vision tasks with the Vision Transformer (ViT) architecture. ViT treats images as sequences of non-overlapping patches, which are then flattened into 1D sequences to be processed by the Transformer. By leveraging self-attention, ViT captures global contextual information from the entire image and allows for efficient modeling of long-range dependencies.

The self-attention mechanism is the cornerstone of the Transformer framework. It computes a weighted sum of values (representations of input elements) based on their relevance to a query (a representation to be updated). Self-attention computes attention scores between each pair of elements in the input sequence and generates attention weights that signify the importance of different elements relative to each other. This attention mechanism allows the model to adaptively focus on relevant input parts during each processing step, facilitating better representation learning [64].

In the Transformer architecture, each self-attention layer is followed by a feed-forward neural network module. The feed-forward module consists of two linear transformations separated by a non-linear activation function, typically a GELU (Gaussian Error Linear Unit) or ReLU (Rectified Linear Unit). This module introduces non-

linearity and enables the model to learn complex relationships between different elements in the input sequence. Combining self-attention and feed-forward modules empowers the Transformer to effectively model local and global dependencies within the input sequence, making it highly adaptable to various tasks, including computer vision [12].

Recently, with the advancement of transformers [64] in computer vision fields [17], [27], [63], several SOTA scene text spotting methods [3], [13], [30], [49], [51], [53] proposed to take the benefit of transformer-based pipelines in their framework. These methods achieved superior performance in both regular and irregular benchmark datasets. For example, Kittenplon et al. [28] utilized a transformer-based detector, Deformable-DETR [70], as its primary framework by proposing a multi-task prediction head that can output word instances and box coordinates of an arbitrary shape text. [68] used transformers as the main block for an end-to-end text-spotting framework for text detection and recognition in wild images. These methods removed the dependency of region-of-interest operations and post-processing stages in their framework. Thus, they can output both Bezier curve and polygon representations and achieve superior benchmark performance. Very recently, Raisi et al. [54] proposed an end-to-end framework for scene text spotting that is also capable of improving the recognition performance for an adverse situation like occlusion. This method utilized an MAE in their pipeline equipped with a powerful detector, namely Deformable-DETR [70], to capture the arbitrary shape of occluded text instances in the wild images. In this study, a pre-trained model from [54] is utilized for the VPR task.

The Proposed Scene Text Spotting Architecture

For complete text reading, simultaneous text detection and recognition are required. Unlike stepwise detection and recognition, as utilized in [22], the end-to-end framework will improve the overall speed by eliminating multiple processing steps. Furthermore, an end-to-end transformer is expected to offer higher accuracy compared to previous end-to-end CNN-based approaches [38], [39].

A. Backbone

The overall framework of the proposed method is shown in Fig. 2 Inspired [32], the proposed model uses pre-trained models of the Vision Transformer architecture (ViT) [12] as the backbone. The 2-Dimensional (2D) input image ($I \epsilon \mathbb{R}^{H\times W\times C}$) is first split into a non-overlapping sequence of patches ($I' \epsilon \mathbb{R}^{N\times P^2\times C}$), where ($H$,$W$) represent the height and width of the image, C is the number of channels and (P, P) denote the resolution of the patches. The number of patches ($N = HW/P^2$) is set to 16.

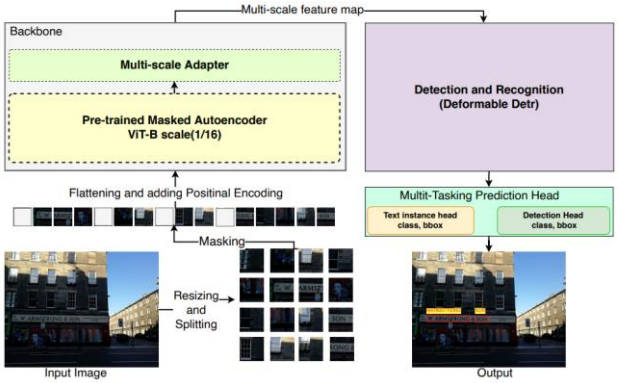

Fig. 2: Block diagram of the proposed scene text spotting architecture using a transformer for VPR [54]. Unlike the stepwise pipeline in [22], the proposed model outputs the bounding box coordinates and the word instances in an end-to-end manner for the VPR task. It is best viewed when zoomed in.

After masking a large set of the input patches ($\sim 75\%$) and adding the 1D position embedding, these patches are passed into the encoder of the MAE ViT block, which contains several multi-head self-attention and feed-forward modules. The encoder operates on unmasked patches to acquire the visual feature embeddings. However, the final output of the ViT encoder backbone is single-scale due to the columnar structure of ViT, which makes them inadequate for detecting multi-scale text instances. To address this, a multi-scale adapter module [32] is utilized. It is worth mentioning that the model uses a pre-trained MAE [18] (ViT-Base/16) as the backbone for feature extraction. This backbone was further fine-tuned on 36 classes of alphanumeric characters (More details in subsection A of Experimental Results).

B. Multi-scale adapter

Inspired by [32], [54], a single-scale ViT into the multi-scale FPN for capturing different resolutions of text regions is adapted. The multi-scale feature map module utilizes the idea of up-sampling or down-sampling into the intermediate single-scale ViT's feature map with a columnar structure [32]. The Multi-scale Adapter module in Fig. 2 consists of 4 up-sampling and down-sampling subblocks. The output feature map of the first block undergoes up-sampling with a scaling factor of 4. Subsequently, the output of the following block is up-sampled, but this time by a factor of 2. On the other hand, the output of the third block remains unchanged, which remains equal to the original feature map. Finally, the output feature map of the last block undergoes down-sampling with a scaling factor of 2. As a result of these operations, a set of multi-scale features is obtained, containing feature maps with different resolutions.

These multi-scale feature maps are then fed into an extended detector [70], which utilizes this information to perform text detection and recognition. By leveraging the diverse scales and resolutions captured in the multi-scale

features, the upgraded detector can effectively handle text instances of varying sizes and efficiently analyze the input image at different levels of detail.

C. Text Predictor

After feature extraction and multi-scaling, the resulting feature maps are fed to the text of the final module to detect and recognize the text instance of a given image. As shown in Fig. 2, the proposed text predictor leverages a modified Deformable-DETR [70] with a multi-task prediction head. This work introduces an enhanced adaptation of the FFN layer, which differs from the [70] architecture. The proposed modifications aim to significantly enhance its ability to capture the distinctive text features produced by the encoder's Multi-Head Self Attention (MHSA) mechanism. The improved FFN now comprises two layers of 1×1 convolutions, supplemented by ReLU activations, and ultimately integrated with a residual connection. This innovative approach effectively amplifies the FFN's capacity to encapsulate and process essential information from the encoder's MHSA, resulting in a more robust and efficient Transformer model.

During training, the encoder's multi-head self-attention detector learns how to separate individual characters and word instances in the scene image by performing global computations. The decoder typically learns how to attend to a different part of characters in words by using different learnable vectors (so-called object queries). During training, the multi-task head (last layer of the decoder) can directly predict both absolute bounding box coordinates and sequence of characters, eliminating the use of any hand-designed components and post-processing like anchor design and non-max suppression. To achieve this, a novel loss function based on optimal bipartite matching between the predicted text instances and the corresponding ground truth is leveraged. This matching process is crucial as it allows us to establish one-to-one correspondences and is efficiently computed using the Hungarian algorithm [70] explicitly adapted for this task. Using the Hungarian algorithm, the model can determine the optimal matching between the predicted and ground-truth elements. This matching information is instrumental in evaluating the performance of the Transformer model for character and word prediction within the text regions. The final loss function takes advantage of these optimal correspondences, enabling the model to learn and improve its predictions more effectively, enhancing accuracy and performance in the task. This work implements the same text filtering criteria introduced in [22] for comparing the query and inference frames.

Experimental Results

A. Implementation Details

The final model is trained on 4 GPUs of NVidia A100. First, by using about 500K cropped alphanumeric synthetic character images from the SynthText dataset [16] for 20 epochs are used to train the pre-trained encoder backbone of MAE (ViT-Base/16) [22] to make it more appropriate for scene text detection and recognition application for 200 epochs. Subsequently, 300 images of ICDAR15 [25] datasets are combined to fine-tune the final model. The Deformable DETR [70] module's object queries are set to 300, and the AdamW optimizer is used to optimize the model's parameters. The following augmentation strategies are applied to the input images during the learning process: horizontal and vertical flip, image resizing, brightness, contrast, and saturation. The model is trained with a batch size of 2 per GPU by employing a learning rate of 1×10^{-4} throughout the training process, and the whole process of training time takes ~ 23 hours. An NVIDIA RTX 3080Ti GPU with 12GB of memory is used for testing the final model.

B. Datasets

The **Self-Collected TextPlace (SCTP)** Dataset [22] is designed explicitly for visual place recognition tasks in urban places. The images of this dataset are captured using a side-looking mobile phone camera. These images include three pairs of map and query sequences in outdoor streets and an indoor shopping mall and contain significant challenging scenarios, including high dynamics, random occlusions, illumination changes, irregular text instances, and viewpoint changes.

The **ICDAR15** [25] is a challenging dataset that contains various indoor and outdoor multi-oriented text instances. Like most of the images in the VPR applications, this dataset has a wide variety of blurry and low-resolution text. The text instances in this dataset are annotated in quadrilateral bounding box annotations.

C. Evaluation Metrics

In the context of text detection and visual place recognition, Precision, Recall, and H-mean are widely used evaluation metrics to assess the performance of the trained model. Precision, recall, and H-mean are measurements that are accepted in almost all text detection communities [23]-[26], [35]-[42], [65]-[67]. These metrics are based on comparing the predicted text regions and the ground truth (manually annotated) text regions in an image, which can be described as follows:

Precision: measures the proportion of predicted text regions correctly identified as text regions among all the predicted text regions. It quantifies the model's ability to avoid false positives. The formula for precision is:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

where True Positives (TP) are the number of correctly predicted text regions, and False Positives (FP) are the number of non-text regions incorrectly predicted as text

regions.

Recall: measures the proportion of correctly predicted text regions out of all the ground truth text regions in the image. It assesses the model's ability to avoid false negatives. It can be defined as follows:

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

where False Negatives (FN) are the number of text regions that the model did not correctly predict.

H-mean (Harmonic Mean): The H-mean, the F1-score, is the harmonic mean of precision and recall. It provides a balanced measure that considers both precision and recall simultaneously. The formula for H-Mean is:

$$H - Mean = 2\frac{Precison \times Recall}{Precison + Recall} \quad (3)$$

These metrics are essential in text detection evaluation [25], as they provide insights into the model's accuracy in correctly identifying text regions and its ability to balance false positives and false negatives. Researchers use these metrics to compare text detection algorithms and fine-tune models for optimal performance.

In this work, to compare the performance of the proposed model with SOTA VPR methods [1], [2], [22] [22], the same evaluation metrics, namely, precision-recall evaluation measurements are followed as in [22], [62].

Fig. 3: Query image and matching reference examples of [22] dataset. The proposed model detects and recognizes the most challenging text instances required to match the frames of the query (top column) and reference (bottom column). It is best viewed in color when zoomed in. The output results are indicated in 'cyan' color.

For end-to-end text detection and recognition that aims to output the correct string of the word instances in the image, in addition to the detected text metrics, the H-mean (F1-score) is used for the evaluation [3], [13], [14], [30], [31], [33], [38], [39], [43], [47], [49], [51], [57].

D. Quantitative Comparison of the VPR SCTP dataset with SOTA methods

The quantitative results of the proposed model with several SOTA methods [1], [2], [11], [22], [45] on the SCTP dataset [22] are shown in Table 1. The proposed model achieved the best performance in terms of recall for this dataset, which contains significant challenges like irregular and partially occluded text instances. This performance confirms the effectiveness of the proposed method for VPR.

Table 1: Precision-Recall comparison of the proposed model with SOTA methods including TextPlace [22], ToDayGAN [1], NetVLAD [2], SeqSLAM [45], and FAB-MAP [11] using SCTP [22] dataset. The best performance is highlighted in bold.

Model	Recall				
	0.2	0.4	0.6	0.8	0.9
Proposed model	1	1	1	**0.97**	**0.93**
TextPlace	1	1	1	0.96	0.91
NetVLAD-10	1	1	1	0.95	**0.93**
NetVLAD-20	1	1	1	0.91	0.87
NetVLAD-30	1	1	0.97	0.85	0.83
ToDayGAN-10	0.50	0.55	0.58	0.57	0.56
ToDayGAN-20	0.40	0.40	0.40	0.38	0.38
ToDayGAN-30	0.26	0.24	0.24	0.25	0.24
FAB-MAP-10	0.79	0.69	0.67	0.65	0.63
FAB-MAP-20	0.76	0.69	0.67	0.63	0.60
FAB-MAP-30	0.68	0.67	0.67	0.62	0.58
SeqSLAM	0.30	0.24	0.18	0.13	0.13

E. Qualitative Comparison on the VPR SCTP dataset

Fig. 3 illustrates the qualitative results of the proposed model on the SCTP [22] dataset. As seen, the model successfully read challenging text instances of both query and reference frames. The results of the proposed model are also compared with some of the SOTA techniques [2], [22], [45], as shown in Fig. 4; the proposed text spotting model correctly matches the query frame with frame in inference.

Fig. 4: Qualitative comparison of the proposed model with SOTA methods [2], [22], [45] on the SCTP dataset. The correct and incorrect results are bounded with green and red colors.

F. Quantitative Comparison with SOTA text detection and recognition approaches using ICDAR Dataset.

The proposed model also is compared with some SOTA scene text detection and recognition approaches [5], [34], [38], [71], [72], [74], [75]. As seen from Table 2, while these methods are trained on many images of synthetic

datasets and fine-tuned on real-world datasets, the proposed model achieves the best performance (P = 90.2) regarding precision for text detection and competitive performances in terms of recall and H-mean. It also performed well while testing for end-to-end text detection and recognition (E2E H-mean = 68.2). Since using the number of training images affects the final performance of deep learning models, for a fair comparison, only SOTA methods that used close images to the proposed model are selected for comparison.

Table 2: Quantitative comparison of the model with the baseline Textboxes++ [34] and other SOTA text detection and recognition methods using the ICDAR15 [25] dataset. P, R, H, and F mean Precision, Recall, H-mean, and F-measure, respectively. E2E denotes end-to-end text spotting, and FPS is Frames per second. The best and second-best performances are highlighted in bold and underlined.

Model	Detection			E2E	FPS
	P	R	H	F	
CRAFT [5]	88.5	**84.69**	**86.9**	--	--
PSENet [74]	86.9	84.5	85.6	--	--
EAST [75]	83.3	78.3	80.7	--	--
FOTS [38]	88.8	82.0	85.3	--	--
DRGN [71]	88.5	84.6	86.5	--	--
CharNetR50 [72]	--	--	--	<u>60.72</u>	--
Textboxes++[34]	87.8	78.5	82.9	51.9	2.3
Proposed Model	**90.2**	<u>83.1</u>	<u>86.5</u>	**68.2**	**11.0**

G. *Ablation Experiments*

1) Output Feature Comparison of the Proposed Model and VPR Methods. As mentioned in the Introduction, VPR algorithms mainly design their architecture to extract Point features, also known as keypoint features, to represent and match distinctive points between images for place recognition tasks. These algorithms detect and describe keypoint features based on local image information around the keypoints. To compare the output semantic features extracted from the proposed model and the keypoint features of VPR techniques, a qualitative ablation study is conducted, which is shown in Fig. 5; the text features of the proposed model are more semantic indexes and fewer in number compared to the keypoints extracted from other VPR approaches.

2) Model Comparison with the Baseline Text Spotting Utilized in the VPR Application. The TextPlace [22] model uses the pre-trained model of Textboxes++ [34] algorithm as the primary text extraction in their framework for the VPR application.

In this section, additional experiments to compare the proposed model with Textboxes++ [34] are conducted and provided quantitative and qualitative results to show how the proposed model performs for text instances that appear in the wild images using the benchmark dataset, ICDAR15 [25], as in [34].

Fig. 5: Comparing the (a) Text features used in the proposed E2E text detection model and (b) keypoint features used in different VPR techniques [11], [45].

Table 2 shows the quantitative comparison of the Textboxes++ that is used as a baseline [22] and the proposed model using the well-known text detection and end-to-end text spotting evaluation metrics [25]. As seen, the proposed approach outperformed the [34] in both detection and end-to-end spotting tasks. It achieves an H-mean detection performance of 86.5% compared to 82.9% in [34]. It also surpasses the Textboxes++ method with a large margin of $\sim 16\%$ in end-to-end F-measure performance. Furthermore, the proposed model is more suitable for real-time detection and recognition as it provides better FPS. These performances confirm the proposed models' good generalization and efficiency on challenging and unseen VPR dataset, SCTP (see Table 1).

To see how the proposed algorithm performs in challenging cases of the ICDAR15 dataset, a qualitative comparison of the proposed model with failure cases in [34] is provided. As shown in Fig.6, the proposed model successfully predicted most of the failure cases. Since text instances in the wild images usually appear with arbitrary shapes, it is important to use a model that better captures any shape of the scene text. The results in the last column in Fig. 6 also show that the proposed pipeline is capable of accurately outputting polygon bounding boxes for curved text instances, whereas Textboxes++ fails to detect.

3) Qualitative results on the challenging text sample images in the wild. To show the proposed model's capability and its limitation on challenging real-world scenarios, more qualitative results by showcasing visual examples of successful and unsuccessful predictions are provided. As shown, the proposed model performed well

on different challenging text instances in the wild images, such as partial occlusion, complex fonts, Illumination variation, oriented text, and curved text.

Fig. 6: Comparison between the proposed end-to-end model (bottom row) with Textboxes++ [34] algorithm (top row). The red boxes in the top row images show the failure cases (Images are taken from [34]), and the cyan text and boxes show our results. The orange arrows point to text regions where the proposed model could successfully predict failure text instances of Textboxes++.

All the images in Fig. 7 are selected from the datasets different from ICDAR15 used during fine-tuning and testing. For example, there is no curved annotation in ICDAR15, but the proposed model could accurately bound a curved bounding box around those text instances (Fig. 7.b, Fig. 7.c, and Fig. 7.d.). The performance and other successful predictions show that the proposed model could be generalizable on unseen sample text instances from the TotalText [76] and CTW1500 [77] datasets designed for irregular text detection and recognition and different challenges than ICDAR15. In addition, another experiment by applying partial occlusion on the characters is conducted. As seen in the letter 'G' in Fig. 7.b, letter 's' in Fig. 7.c, letter 'c,' 'k,' and letter 'r' in Fig. 7.d., the proposed model correctly recognizes those letters, confirming its capability of partial occlusion detection and recognition due to the capability of individual character spotting and masking the input images of the proposed architecture. However, from Fig. 7, the proposed model performed poorly on low-resolution and low-contrast text instances.

4) Inference Speed of the model. This work also experiments with the inference speed of the proposed model and compares it with [22] in terms of Frames Per Second (FPS). To that effect and for a fair comparison, an RTX 3080Ti GPU is used that has a similar memory used in [22] and presented in [34]. The proposed model outperformed the TextPlace method by a large margin, achieving a $\sim$ 11 FPS in compared to 2.3 FPS in [22].

G. Limitations and Future Work

As mentioned in the previous sections, although the proposed model performed well on many challenging cases in the wild images, there are many shortcomings that can be improved or addressed in future work. First, the model needed character annotation to be trained. These types of annotations are expensive to prepare. To address this problem, weakly supervised or unsupervised learning techniques can be applied to the model.

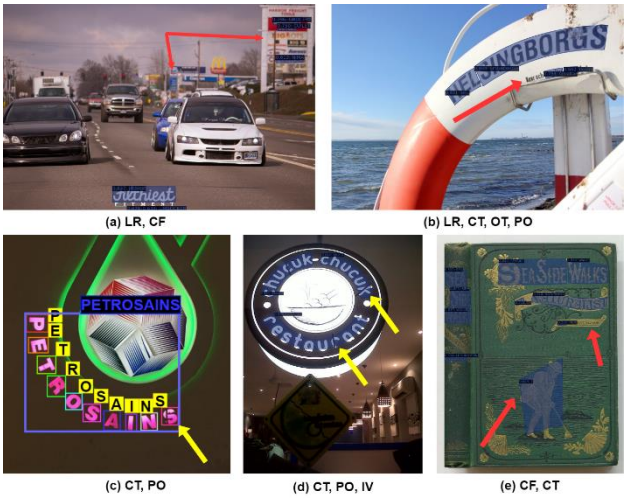

Fig. 7: Qualitative results of the proposed model on some challenging examples images, where PO: Partial Occlusion, CF: Complex Fonts, IV: Illumination Variation, LR: Low Resolution, OT: Oriented Text, and CT: Curved Text. The red and yellow arrows represent failure cases and the partially occluded characters, respectively. The above images are selected from two benchmarks: TotalText and CTW100 datasets designed for curved text detection and recognition in the wild. The proposed model needs to be trained in these images. The accurate detection of the text instances shows its generalizability on unseen images.

The proposed model performs poorly on low-resolution, blurry, and high-occluded text instances, as seen in Fig. 7. These challenges are still open in many SOTA text detection and recognition in the wild methods, and humans may need help reading these types of text instances. However, one way to address these problems is to benefit from the recent advancement in natural language processing algorithms like combining the pre-trained language model modules like Generative Pre-training Transformer (GPT) [78] and compositionality techniques [79] in the text detection and recognition framework to help the model to guess the uncaptured characters in the text.

Conclusion

In this work, an end-to-end scene text spotting model for the visual place recognition task is presented. The proposed model has leveraged a robust SOTA backbone of pre-trained MAE and a modified multi-task transformer detector. The quantitative and qualitative experimental results have shown that the proposed model outperforms SOTA models in VPR, which confirms the robustness of the proposed end-to-end scene text detection and recognition model. It obtained the best performance in

terms of precision-recall for the benchmark VPR application dataset, called SCTP. The proposed model outperformed the baseline text detection and recognition technique, textboxes++, used in TextPlace by a large margin regarding precision, recall, and H-mean. It also achieved competitive performance with many SOTA text detection and recognition techniques. The qualitative ablation experiments also confirmed that the proposed model could spot many challenging text instances in the wild images, including rotated and curved, complex fonts, partial illumination variation, and occlusion. The limitations and future work to improve the performance of the proposed model are also discussed. Other applications besides VPR include different facets of localization and mapping. Detecting and recognizing text allows the potential to leverage semantics and the features related to the detected text to localize better and map instead of just using indirect features.

Acknowledgment

We would like to thank the Ontario Centers of Excellence (OCE), the Natural Sciences and Engineering Research Council of Canada (NSERC), and ATS Automation Tooling Systems Inc., Cambridge, ON, Canada, for supporting this research work.

Author Contributions

Z. Raisi collected the data, implemented the code, carried out the analysis, and wrote paper. Dr. J. Zelek interpreted the results and supervised the research.

Conflict of Interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication or falsification, double publication and, or submission, and redundancy, have been completely witnessed by the authors.

Abbreviations

ViT	Vision Transformer
FPS	Frames per Second
SCTP	Self-Collected Text Place
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
DETR	Detection using Transformers
SOTA	State Of the Art
VPR	Visual Place Recognition
MAE	Masked Autoencoders
SLAM	Simultaneous Localization and Mapping

References

[1] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, L. Van Gool, "Night-to-day image translation for retrieval-based localization," in Proc. 2019 International Conference on Robotics and Automation (ICRA): 5958–5964, 2019.

[2] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in Proc. IEEE/CVF International Conference on Computer Vision: 5297–5307, 2016.

[3] R. Atienza, "Vision transformer for fast and efficient scene text recognition," Document Analysis and Recognition – ICDAR 2021. Springer International Publishing, pp. 319–334, 2021.

[4] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, H. Lee, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in Proc. International Conference on Computer Vision (ICCV), 2019.

[5] Y. Baek, B. Lee, D. Han, S. Yun, H. Lee, "Character region awareness for text detection," in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2019.

[6] Y. Baek, S. Shin, J. Baek, S. Park, J. Lee, D. Nam, H. Lee, "Character region attention for text spotting," ArXiv, vol. abs/2007.09629, 2020.

[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers," arXiv preprint arXiv:2005.12872, 2020.

[8] W. Chan, C. Saharia, G. Hinton, M. Norouzi, N. Jaitly, "Imputer: Sequence modeling via imputation and dynamic programming," arXiv preprint arXiv:2002.08926, 2020.

[9] C. K. Ch'ng, C. S. Chan, "Total-text: A comprehensive dataset for scene text detection and recognition," in Proc. IAPR International Conference on Document Anal. and Recognition (ICDAR), 1: 935–942, 2017.

[10] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al., "Rethinking attention with performers," arXiv preprint arXiv:2009.14794, 2020.

[11] M. Cummins, P. Newman, "Fab-map: Probabilistic localization and mapping in the space of appearance," Int. J. Rob. Res., 27(6): 647–665, 2008.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[13] S. Fang, H. Xie, Y. Wang, Z. Mao, Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition: 7098–7107, 2021.

[14] W. Feng, W. He, F. Yin, X. Y. Zhang, C. L. Liu, "Textdragon: An end-to-end framework for arbitrarily shaped text spotting," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition: 9076–9085, 2019.

[15] S. Garg, T. Fischer, M. Milford, "Where is your place, visual place recognition?" arXiv preprint arXiv:2103.06443, 2021.

[16] A. Gupta, A. Vedaldi, A. Zisserman, "Synthetic data for text localization in natural images," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 2315–2324, 2016.

[17] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, et al., "A survey on the visual transformer," arXiv preprint arXiv:2012.12556, 2020.

[18] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, R. Girshick, "Masked autoencoders are scalable vision learners," arXiv preprint arXiv:2111.06377, 2021.

[19] K. He, G. Gkioxari, P. Dollar, R. Girshick, "Mask R-CNN, " in Proc. IEEE International Conference on Computer Vision: 2961–2969, 2017.

[20] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 770–778, 2015.

[21] S. Hochreiter, J. Schmidhuber, "Long short-term memory," Neural Comput., 9(8): 1735–1780, 1997.

[22] Z. Hong, Y. Petillot, D. Lane, Y. Miao, S. Wang, "Textplace: Visual place recognition and topological localization through reading scene texts," in Proc. IEEE/CVF International Conference on Computer Vision: 2861–2870, 2019.

[23] M. Iwamura, N. Morimoto, K. Tainaka, D. Bazazian, L. Gomez, D. Karatzas, "ICDAR2017 robust reading challenge on omnidirectional video," in Proc. 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 1: 1448–1453, 2017.

[24] M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition, " arXiv preprint arXiv:1406.2227, 2014.

[25] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, et al., "ICDAR 2015 competition on robust reading," in Proc. International Conference on Document Analysis and Recognition (ICDAR): 1156–1160, 2015.

[26] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. I Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. De Las Heras, "ICDAR 2013 robust reading competition," in Proc. International Conference on Document Analysis and Recognition: 1484–1493, 2013.

[27] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, M. Shah, "Transformers in vision: A survey," arXiv preprint arXiv:2101.01169, 2021.

[28] Y. Kittenplon, I. Lavi, S. Fogel, Y. Bar, R. Manmatha, P. Perona, "Towards weakly-supervised text spotting using a multi-task transformer," arXiv preprint arXiv:2202.05508, 2022.

[29] A. B. Laguna, K. Mikolajczyk, "Key. net: Keypoint detection by handcrafted and learned CNN filters revisited," IEEE Trans. Pattern Anal. Mach. Intell., 45(1): 698-711, 2022.

[30] J. Lee, S. Park, J. Baek, S. Joon Oh, S. Kim, H. Lee, "On recognizing texts of arbitrary shapes with 2D self-attention," in Proc. IEEE CVPR: 546–547, 2020.

[31] H. Li, P. Wang, C. Shen, "Towards end-to-end text spotting with convolutional recurrent neural networks," in Proc. 2017 IEEE International Conference on Computer Vision (ICCV): 5248–5256, 2017.

[32] Y. Li, S. Xie, X. Chen, P. Dollar, K. He, R. Girshick, "Bench-marking detection transfer learning with vision transformers," arXiv preprint arXiv:2111.11429, 2021.

[33] M. Liao, G. Pang, J. Huang, T. Hassner, X. Bai, "Mask textspotter v3: Segmentation proposal network for robust scene text spotting," in Proc. Computer Vision–ECCV 2020: 16th European Conference, Part XI 16: 706–722, 2020.

[34] M. Liao, B. Shi, X. Bai, "Textboxes++: A single-shot oriented scene text detector," IEEE Trans. Image Process., 27(8): 3676–3690, 2018.

[35] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C. L. Zitnick, "Microsoft coco: Common objects in context," in Proc Eur. Conference on Computer Vision. Springer: 740–755, 2014.

[36] V. Nazarzehi, R. Damani, "Decentralised optimal deployment of mobile underwater sensors for covering layers of the ocean," Indones. J. Electr. Eng. Comput. Sci., 25(2): 840–846, 2022.

[37] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, "SSD: Single shot multibox detector," in Proc. Eur. Conference on Computer Vision. Springer: 21–37, 2016.

[38] X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, J. Yan, "FOTS: Fast oriented text spotting with a unified network," in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 5676–5685, 2018.

[39] Y. Liu, H. Chen, C. Shen, T. He, L. Jin, L. Wang, "Abcnet: Real-time scene text spotting with adaptive bezier-curve network," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition: 9809–9818, 2020.

[40] Y. Liu, C. Shen, L. Jin, T. He, P. Chen, C. Liu, H. Chen, "Abcnet v2: Adaptive bezier-curve network for real-time end-to-e end text spotting," arXiv preprint arXiv:2105.03620, 2021.

[41] S. Lowry, N. S. Underhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, M. J. Milford, "Visual place recognition: A survey," IEEE Trans. Rob., 32(1): 1–19, 2015.

[42] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, "ICDAR 2003 robust reading competitions," in Proc. Seventh Int. Conference on Document Analysis and Recognition: 682– 687, 2023.

[43] P. Lyu, M. Liao, C. Yao, W. Wu, X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in Proc. Eur. Conference on Computer Vision (ECCV) : 67–83, 2018.

[44] C. Masone, B. Caputo, "A survey on deep visual place recognition," IEEE Access, 9: 19516–19547, 2021.

[45] M. J. Milford, G. F. Wyeth, "Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights," in Proc. IEEE International Conference on Robotics and Automation: 1643–1649, 2012.

[46] A. Mishra, K. Alahari, C. V. Jawahar, "Scene text recognition using higher order language priors," in BMVC, 2012.

[47] S. Qin, A. Bissacco, M. Raptis, Y. Fujii, Y. Xiao, "Towards unconstrained end-to-end text spotting," in Proc. IEEE/CVF International Conference on Computer Vision: 4704–4714, 2019.

[48] T. Q. Phan, P. Shivakumara, S. Tian, C. Lim Tan, "Recognizing text with perspective distortion in natural scenes," in Proc. IEEE International Conference on Computer Vision: 569–576, 2013.

[49] Z. Raisi, M. Naiel, P. Fieguth, S. Wardell, J. Zelek, "2d positional embedding-based transformer for scene text recognition," J. Comput. Vision Imaging Syst., 6(1): 1–4, 2021.

[50] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, J. Zelek, "Text detection and recognition in the wild: A review," arXiv preprint arXiv:2006.04305, 2020.

[51] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell, J. Zelek, "2lspe: 2d learnable sinusoidal positional encoding using a transformer for scene text recognition," in Proc. Conference on Robots and Vision (CRV): 119–126, 2021.

[52] Z. Raisi, M. A. Naiel, G. Younes, S. Wardell, J. S. Zelek, "Transformer-based text detection in the wild," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops: 3162–3171, 2021.

[53] Z. Raisi, G. Younes, J. Zelek, "Arbitrary shape text detection using transformers," in Proc. IEEE International Conference on Pattern Recognition (ICPR): 3238-3245, 2022.

[54] Z. Raisi, J. Zelek, “Occluded text detection and recognition in the wild,” in IEEE Proceeding Conference on Robots and Vision (CRV): 140-150, 2022.

[55] Z. Raisi, J. S. Zelek, “End-to-end scene text spotting at character level,” J. Comput. Vision Imaging Syst., 7(1): 25-27, 2021.

[56] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 779–788, 2016.

[57] S. Ren, K. He, R. Girshick, J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. in Neural Info. Process. Syst.: 91–99, 2015.

[58] A. Risnumawan, P. Shivakumara, C. S. Chan, C. L. Tan, “A robust arbitrary text detection system for natural scene images,” Expert Syst. Appl., 41(18): 8027–8048, 2014.

[59] D. E. Rumelhart, G. E. Hinton, R. J. Williams, “Learning representations by back-propagating errors,” Nature, 323(6088): 533–536, 1986.

[60] A. Shahab, F. Shafait, A. Dengel, “ICDAR 2011 robust reading competition challenge 2: Reading text in scene images,” in Proc. International Conference on Doc. Anal. and Recognition: 1491–1496, 2011.

[61] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, X. Bai, “Aster: An attentional scene text recognizer with flexible rectification,” IEEE Trans. Pattern Anal. Mach. Intell., 41(9): 2035-2048, 2018.

[62] Y. Sun, Z. Ni, C.-K. Chng, Y. Liu, C. Luo, C. C. Ng, J. Han, E. Ding, J. Liu, D. Karatzas, et al., “ICDAR 2019 competition on large-scale street view text with partial labeling –RRC-LSVT,” arXiv preprint arXiv:1909.07741, 2019.

[63] Y. Tay, M. Dehghani, D. Bahri, D. Metzler, “Efficient transform- ers: A survey,” arXiv preprint arXiv:2009.06732, 2020.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, “Attention is all you need,” in Proc. Advances in Neural Information Processing Systems (NIPS 2017): 5998– 6008, 2017.

[65] K. Wang, S. Belongie, “Word spotting in the wild,” in Proc. Eur. Conference on Computer Vision. Springer: 591–604, 2010.

[66] C. Yao, X. Bai, W. Liu, Y. Ma, Z. Tu, “Detecting texts of arbitrary orientations in natural images,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition: 1083–1090, 2012.

[67] L. Yuliang, J. Lianwen, Z. Shuaitao, Z. Sheng, “Detecting curve text in the wild: New dataset and new solution,” in arXiv preprint arXiv:1712.02170, 2017.

[68] X. Zhang, Y. Su, S. Tripathi, Z. Tu, “Text spotting transformers,” arXiv preprint arXiv:2204.01918, 2022.

[69] X. Zhang, L. Wang, Y. Su, “Visual place recognition: A survey from deep learning perspective,” Pattern Recognit., 113: 107760, 2021.

[70] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, “Deformable detr: Deformable transformers for end-to-end object detection,” arXiv preprint arXiv:2010.04159, 2020.

[71] S. X. Zhang, X. Zhu, J. B. Hou, C. Liu, C. Yang, H. Wang, X. C. Yin, “Deep relational reasoning graph network for arbitrary shape text detection,” in Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition: 9699-9708, 2020.

[72] L. Xing, Z. Tian, W. Huang, M. R. Scott, “Convolutional character networks,” in Proc. the IEEE/CVF International Conference on Computer Vision: 9126-9136, 2019.

[73] I. Loshchilov, F. Hutter, “Decoupled weight decay regularization,” in Proc. International Conference on Learning Representations, 2018.

[74] G. Liao, Z. Zhu, Y. Bai, T. Liu, Z. Xie, “PSENet-based efficient scene text detection,” EURASIP J. Adv. Signal Process., 97(1), 1-13, 2021.

[75] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou,W. He, J. Liang, “East: an efficient and accurate scene text detector,” in Proc. the IEEE Conference on Computer Vision and Pattern Recognition: 5551-5560, 2017.

[76] C. K. Ch'ng, C. S. Chan, “TotalText: A comprehensive dataset for scene text detection and recognition,” in Proc. 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 1: 935-942, 2017.

[77] L. Yuliang, J. Lianwen, Z. Shuaitao, Z. Sheng, “Detecting curve text in the wild: New dataset and new solution,” in arXiv preprint arXiv:1712.02170, 2017.

[78] D. M. Katz, M. J. Bommarito, S. Gao, P. Arredondo, Gpt-4 passes the bar exam. Available at SSRN 4389233.

[79] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, “Building machines that learn and think like people,” Behav. Brain sci., 40, 2017.

Biographies

Zobeir Raisi was born in Chabahar, Iran in 1987. He received his Ph.D. degree in 2022 from the Vision Image Processing Lab (VIPLab) at the Systems Design Engineering Department, University of Waterloo, Waterloo, Ontario, Canada. Currently, he is an assistant professor in the Department of Electrical Engineering at Chabahar Maritime University, Iran. His research interests include computer vision, artificial intelligence, robotics, and image processing.

- Email: zraisi@uwaterloo.ca
- ORCID: 0000-0002-1591-4492
- Web of Science Researcher ID: GLV-1410-2022
- Scopus Author ID: 54897975500
- Homepage: https://uwaterloo.ca/scholar/zraisi

John Zelek received his Ph.D. degree in Philosophy of Electrical Engineering from the Centre for Intelligent Machines (CIM), McGill University, Montreal, QC, Canada, in 1996. He is currently a Professor of the Systems Design Engineering Department, University of Waterloo, Waterloo, ON, Canada, and the Co-Director of the Vision Image Processing (VIP) Laboratory, University of Waterloo, ON, Canada. He has published over 300 refereed articles, has been a co-founder of five different startup companies from the University of Waterloo, and has been an advisor for various other companies. His research interests include computer vision, AI, robotics, infrastructure monitoring, autonomous vehicles, image processing, augmented reality, and assistive technology, to name a few.

- Email: jzelek@uwaterloo.ca
- ORCID: 0000-0002-8138-3546
- Web of Science Researcher ID: NA
- Scopus Author ID: 6603746225

Homepage: https://uwaterloo.ca/systems-design-engineering/profile/jzelek

How to cite this paper:
Z. Raisi, J. Zelek, "Text detection and recognition for robot localization" J. Electr. Comput. Eng. Innovations, 12(1): 163-174, 2024.
DOI: 10.22061/jecei.2023.9857.658
URL: https://jecei.sru.ac.ir/article_1986.html