TraffSign: Multilingual Traffic Signboard Text Detection and Recognition for Urdu and English

Muhammad Atif Butt[1(✉)], Adnan Ul-Hasan[2], and Faisal Shafait[1,2]

[1] School of Electrical Engineering and Computer Science (SEECS), National University of Science and Technology (NUST), Islamabad, Pakistan
matifbutt@outlook.com

[2] Deep Learning Laboratory, National Center of Artificial Intelligence (NCAI), Islamabad, Pakistan
{adnan.ulhassan,faisal.shafait}@seecs.edu.pk

Abstract. Scene-text detection and recognition methods have demonstrated remarkable performance on standard benchmark datasets. These methods can be utilized in human-driven/self-driving cars to perform navigation assistance through traffic signboard text detection and recognition. Existing datasets include scripts of numerous languages like English, Chinese, French, Arabic, German, etc. However, traffic navigation signboards in Pakistan and many states of India are written in Urdu along with the English translation to guide human drivers. To this end, we present Deep Learning Laboratory's Traffic Signboards Dataset (DLL-TraffSiD) to develop multi-lingual text detection and recognition methods for traffic signboards. In addition, we present a pipeline for multi-lingual text detection and recognition for an outdoor road environment. The results show that our presented system signified better applicability in text-detection and text recognition, and achieved 89% and 92.18% accuracy on the proposed dataset (The proposed dataset along with implementation is available at https://github.com/aatiibutt/TraffSign/).

Keywords: Multi-lingual (Urdu and English) Traffic Signboards Data-set · Multi-lingual text detection · Multi-lingual text recognition

1 Introduction

Traffic signboards are considered as the main source of guidance for human drivers to navigate on the roads. Various types of traffic signboards are placed along highways and roads displaying warnings, speed limits, and directions (written mostly in native languages) to assist human drivers [7]. With the advancement of deep learning, scene-text recognition has drawn a significant attention of the computer vision community, which is evident by large-scale datasets along with scene-text detection and recognition methods [18]. Recently, RoadText-1K [23], COCO-Text

S. Uchida et al. (Eds.): DAS 2022, LNCS 13237, pp. 741–755, 2022.
https://doi.org/10.1007/978-3-031-06555-2_50

[29], and Total-Text [13] datasets consisting of English, Arabic, German, French, Italian, Japanese, and Korean language scripts have been proposed to perform text detection and recognition tasks. However, it is observed from the literature that no significant contribution is made in developing scene text detection and recognition dataset consisting of traffic signboards in Urdu language.

There has been some progress in past couple of years to collect traffic signboard data in Urdu language., Chandio et al. [8] and Arafat and Iqbal [3] have contributed Urdu text detection and recognition datasets; however, one [3] is based on synthetic data where Urdu words are typed on scenic background images whose performance can be influenced on Urdu-text unseen outdoor natural text images. The second dataset [8] consists of text written on random shop boards and wall-choking in non-symmetric font styles and variable text sizes. On the contrary, Urdu-text written on traffic signboards is purely written in sharp-cursive Nastaleeq and Naskh scripts in symmetrical font size and appropriate alignment. In addition, traffic signboards placed on the roads in Pakistan and many states of India (including west Bengal, Uttar Pradesh, Bihar, Jharkhand, etc.) are in Urdu and English languages [3]. These limitations are an indicative of a need for a multi-lingual scene text detection and recognition dataset of traffic signboards along with multi-lingual scene text detection and recognition methods.

To address the above-discussed shortcomings, we have made the following contributions in this research work.

- A multi-lingual (English and Urdu) scene-text detection and recognition dataset - DLL-TraffSiD - has been proposed. This dataset consists of 2,600 text-enriched traffic signboard images such as directional boards, distance boards, instructions, and warning boards, speed limit boards, and location boards with 9,051 bounding boxes annotations containing 13,481 words.
- We have also proposed a scene-text detection and recognition pipeline to perform multi-lingual traffic signboard text detection and recognition.

The remainder of this paper is comprised of the following sections. Section 2 summarizes the related work for scene text detection and recognition and benchmark scene text datasets. Section 3 describes the different steps of developing a DLL-TraffSiD dataset. Section 4 explains the employed methodology for TraffSign and Sect. 5 delineates the experiments and results. Section 6 concludes the paper with some future directions.

2 Related Work

Scene-text understanding methods generally consist of two step processes: (i) text detection, and (ii) text recognition, which are briefly discussed in the following subsections.

2.1 Text Detection

In the early era of ML, researchers introduced handcrafted features-based scene text detection methods. For instance, Hossain et al. [15] proposed Maximally

Stable Extremal Regions (MSER) based road sign text detection from scenic images. Similarly, Basavaraju et al. [5] presented a Laplacian-component analysis fusion-based approach for multilingual text detection from image/video data. However, text recognition under different luminous conditions is still a challenging task. In this regard, Tian et al. [28] presented a Co-occurrence Histogram (Co-HOG) of oriented gradient and convolutional Co-HOG feature descriptor-based scene text detection. Co-HOG is employed to encode context-aware spatial information in neighbor pixels to detect text regions from an input image. However, these methods require a sequence of steps of manual feature extraction along with the optimization in the classifiers that make these methods structurally complex and inefficient.

With the advancement in computer vision, end-to-end representation learning based text detection methods have been proposed. In this regard, Panhwar et al. [21] proposed an Artificial Neural Network (ANN) based text detection method to detect text regions in signboards. Liao et al. [17] presented a Differentiable Binarization (DB) based segmentation network for text localization in scene images. In another research work, Seha et al. [24] introduced an end-to-end Stroke Width Transform (SWT) and Maximally Stable Extremal Regions (MSER) based text detection method in an outdoor environment. However, these methods only focused on unilingual text detection that can be employed in a very limited scale application. To address these issues, Chandio and Pickering [9] proposed a multilingual text-detection method by combining the convolutional layers and the VGG network to perform text detection on Urdu and Arabic text scripts.

2.2 Text Recognition

Convolutional neural networks have demonstrated better applicability in scene-text recognition in various scripts. Bains et al. [4] proposed a text recognition method to recognize Gurumukhi text regions in signboard images. Aberdam et al. [1] presented a Sequence-to-sequence based Contrast Learning method (SeqCLR) for text recognition using the attention-driven technique. However, one main limitation of attention-driven text recognition methods is that the performance can be easily influenced through minor attention drift. To cater with this issue, Cheng et al. [12] proposed a ResNet-based Focusing Attention Network (FAN) to perform scene text recognition tasks in an outdoor environment. Similarly, Lu et al. [19] proposed a multi aspect transformer-based network to perform scene text recognition.

Though deep neural networks can learn robust representations of image artifacts and text style changes, they still run into challenges while coping with scene texts having a pattern and curvature distortions. To cater with this limitation, Shi et al. [25] proposed an end-to-end image-based sequence recognition method for scene text recognition. Chen [11] proposed transformation, attention-driven, and rectification-based networks to improve the text recognition performance in natural scene images. In another research work, Arafat and Iqbal [3] have proposed customized Faster-RCNN based scene text detection, recognition, and ori-

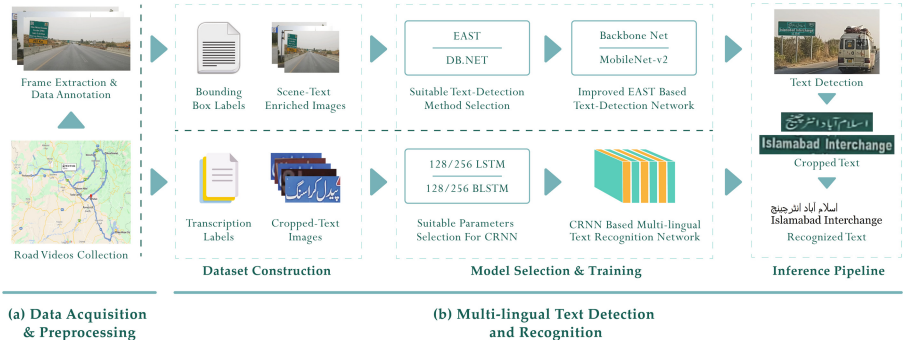


Fig. 1. The Method: Firstly, 900-KM road video sequences are captured, and text-enriched signboard frames are extracted. Secondly, bounding box and transcription annotations are generated to form datasets. In the next step, scene text detection and recognition methods are fine-tuned on proposed datasets. Based on the analysis, best-performing methods are further improved to achieve maximum performance.

entation prediction of Urdu ligatures in outdoor images. The authors employed customized regression residual neural networks for the orientation prediction of text ligatures. Busta et al. [6] proposed a fully connected network-based end-to-end method for multi-language text detection and recognition.

2.3 Standard Benchmark Datasets for Text Detection and Recognition

In recent years, numerous datasets are presented to pave the way towards the generalization of text detection and recognition methods. For instance, Karatzas et al. [16] presented ICDAR 2015 benchmark for incidental scene text detection comprising of 1,600 images. However, the images are captured using google glasses without considering the importance of image quality. Therefore, the models trained over such data get exposed to unseen data vulnerabilities which influence the performance of models in real-world applications. Sun et al. [27] presented Chinese Street View Text (C-SVT) consisting of two chunks; one is completely labeled including bounding boxes of words and characters, while the other consists of annotations of dominant text instances only. Yuan et al. [30] presented Chinese Text in the Wild (CTW) dataset comprising of 32,285 images along with transcription-based annotations. Chandio et al. [10] developed Urdu characters-based dataset containing outdoor scene text images. Gupta et al. [14] presented a synthetic-Synth90K dataset which contains 9 million synthetic images constructed via synthetic text generation engine. Ali et al. [2] presented an Urdu characters-based dataset for the cursive Urdu text recognition in natural outdoor scene images. Shi et al. [26] have presented an ICDAR2017 dataset for text localization and recognition tasks. Nayef et al. [20] introduced ICDAR2019-MLT datasets for multi-lingual scene text detection in an outdoor environment.

Fig. 2. Sample dataset images: Demonstrating the diversity of signboards such as, directional boards, distance boards, location boards, and warning boards.

3 Dataset Preparation

We develop a large-scale text-detection and recognition dataset—DLL-TraffSiD, to overcome the shortcomings in existing text detection and recognition methods, as discussed in Sect. 1. It is worth mentioning that collecting large-scale data through driving videos postures numerous challenges including capturing device configuration, data filtration, and suitable frame selection, ground truth generation alongside the development of multi-lingual text-detection and recognition methods. To accomplish these gigantic tasks, we carefully designed and followed the pipeline, as shown in Fig. 1.

3.1 Data Acquisition and Pre-processing

Driving Platform Setup. A high resolution cameras—GoPro Hero 8, mounted over the dashboard of a standard vehicle to capture the front field of view, is used for data acquisition. The installed camera is configured to 4K resolution with a 16:9 super-wide aspect ratio to capture the ultimate width of the road from the driver's perspective. It is important to mention that the main reason for considering monocular vision over stereo vision is because traffic signboards are placed straight oriented towards the driving road direction [22]. Moreover, the installed camera is stabilized by using a mounting device to avoid vibration effects of the vehicle at varying acceleration and deceleration speed patterns.

Video Sequence Collection. Keeping the limitations of existing datasets in view, we collected driving video sequences of 928 KM covering general traffic roads, highways, and motorways of Khyber-Pakhtunkhwa (KP), Punjab, and Azad Jammu and Kashmir (AJK), Pakistan. We primarily focused on diverse direction boards, warning and distance boards, navigation boards placed alongside the roads, as shown in Fig. 2.

Fig. 3. Dataset-Complexities: Sample images demonstrating inter-ligature and intra-ligature overlapping in Urdu-text instances.

3.2 Multi-lingual Text Detection and Recognition

Dataset Construction. After completing the data collection, the most important task is to align the video sequences to extract frames with signboards and to exclude non-relevant frames. Firstly, we used a frame extraction tool to extract the frames at 30 fps from the video sequences. In the next step, the frames with signboards were manually selected to get a representative subset of signboards considering the diverse text font size, complex scenic backgrounds, writing styles, aspect ratios, and context-sensitivity. Consequently, a subset of 2,600 images is formed as shown in Fig. 2.

Dataset Complexities: Unlike English script, Urdu is a bidirectional cursive script, which makes text detection and recognition tasks more challenging in scene images. In addition, Urdu text is written from right to left in the Nastaleeq script, which is considerably different than the Naksh script that is primarily used to write Arabic. Therefore, detection and recognition of such text become more challenging due to non-uniform inter-ligature overlaps between the letters of the same or two ligatures, as shown in Fig. 3. Moreover, Urdu scripts also include joining and non-joining letters. The joining letters change their shape while merging into the sequence of alphabets within a word, as shown in Fig. 4.

Text Detection Dataset: The proposed dataset contains 2,600 images of traffic signboards. The initial step in annotating text data is to detect the text regions in each frame. In this regard, all the text regions are manually annotated with an enclosed bounding box and four-cornered polygons. The reason for annotating the text regions with polygons is to annotate the tilted text regions accurately. In addition, the legible text regions i.e., single word and illegible text regions i.e., sequences are annotated with one bounding box. Resultantly, 12,179 bounding boxes are generated to perform multi-lingual text detection as shown in Table 1.

Fig. 4. Types of ligatures in Urdu words: Urdu-text instances with single-ligature, two-ligature, and multi-ligature words.

Text Recognition Dataset. We constructed a text recognition dataset containing 9,120 images along with the corresponding multi-lingual transcriptions. It is worth mentioning that the text regions have been cropped from the original images and the corresponding annotations have been generated in UTF-8 encoded text files. All the words/sequences of words are manually annotated and verified for text recognition purposes. In addition to annotations files, two lexicon text files consisting of 50,000 and 80,000 common Urdu and English words are created to address the contextual error while word recognition.

4 The Methodology

This section describes three sub-parts of the TraffSign system. As described in the pipeline, our proposed system includes (i) a multi-lingual text detection network and (ii) a multi-lingual text recognition network. Each of these architectures is briefly elaborated in the below subsections.

4.1 Multi-lingual Text Detection Architecture

The size of word regions varies in the dataset images; therefore, it is important to extract the features from the late stage of the feature extraction block. To choose suitable network for multi-lingual text detection, we fine-tuned existing text detectors, namely, Efficient and Accurate Scene Text (EAST) [31] and Differentiable Binarization (DB) [17] networks.

Table 1. Statistics of the bounding boxes, words, and characters for each script in the DLL-TraffSiD text detection and recognition dataset

Script type	No. of bounding boxes	No. of words	No. of characters
English	3,269	5,828	11,754
Urdu	5,782	7,653	38,521

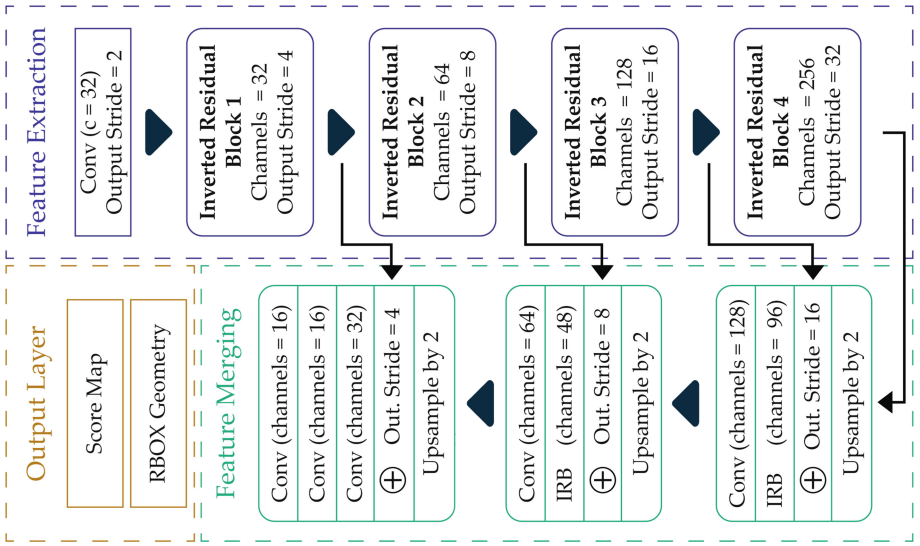


Fig. 5. Improved EAST Architecture with MobileNet-v2 as Backbone for Multi-lingual Text Detection. MobileNet-v2, consisting of four residual blocks is used to perform feature extraction. Each residual block is concatenated with EAST network in feature merging block. Lastly, output layer is used to generate bounding box and score map.

Based on the better performance, EAST architecture has been further improved as shown in Fig. 5. The presented architecture consists of (i) feature extraction, (ii) feature merging, and (iii) output blocks. In the existing EAST based text detection implementations, PA-net, VGG-16, and ResNet-50 architectures [18,31] have been used as feature extraction networks. However, the EAST detector with these feature extraction networks has not performed well in terms of accuracy and precision on our proposed dataset.

To improve upon this shortcoming, we present an improved MobileNet-v2 feature extraction network comprising of four inverted residual blocks including convolution, batch normalization, and pooling layers. Initially, convolution is performed on an input image and the resultant feature map is passed towards the inverted residual blocks. The inverted residual blocks process the feature maps in a feed-forward fashion while also passing the feature merging blocks to concatenate multi-layered information.

In the next step, feature merging block comprising of three blocks connected in a bottom-up fashion. The output feature map of the feature extraction network is up-sampled by a factor of 2 and a series of convolutions are performed with various output numbers of channels over the concatenated feature map. Consequently, the final feature map is passed to the output block to produce the predictions. The output block contains 1×1 convolution projecting 32 channels to generate the score map along with the geometric output containing coordinate-based information of the corresponding text regions.

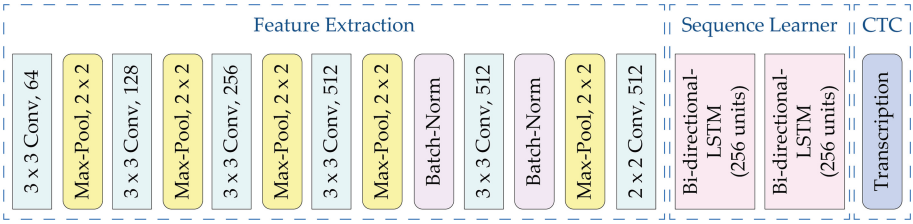


Fig. 6. Bi-Directional LSTM based CRNN Architecture For Multi-lingual Text Recognition. Feature Extraction block contains custom convolution layers followed by max-pooling and activation layers. BLSTMs with 256 units, followed by CTC layer are employed to perform sequence learning and transcription generation.

4.2 Text Recognition Architecture

Multi-lingual text recognition in scenic image data is a challenging task due to variable font size, style, and orientation. To this end, firstly, a base convolutional recurrent neural network (CRNN) has been selected to perform multi-lingual text recognition using both unidirectional and bi-directional Long Short-term Memory (LSTM) with 128 and 256 units, respectively. CRNN is a combination of two categorial neural networks i.e., convolutional and recurrent neural networks. CNN layers in CRNN perform feature extraction from the visual input, whereas RNN layers perform sequence classification. Based on the performance, we further optimized bi-directional LSTM (256-units) based CRNN by inserting batch-normalization followed by max-pooling layer prior to base convolution layer in feature extraction block to achieve maximum performance. As depicted in Fig. 6, a feature extraction block consisting of six convolution layers followed by max-pooling and batch normalization layers is used to extract useful information from the text regions. In the next step, two LSTMs with unidirectional and bi-directional units are used in the architecture to evaluate the text recognition performance on our proposed dataset. Based on better performance, two bi-directional LSTMs with 256 hidden units are used to perform sequence classification. Lastly, Connectionist Temporal Classification (CTC) layer is used to generate final transcription along with the associated confidence score of prediction.

5 Experiments and Results

The experiments of proposed scene text detection and recognition methods are performed on the high-performance machine equipped with the core i9 - 9900k CPU with 32GB RAM, and RTX 2080TI, 11GB DDR5 GPU, having a 64bit windows 10 operating system. Pytorch 1.8.0 library is utilized for data preprocessing, training, and evaluation of the proposed methods. The implementation guide along with training and evaluation codes are available at https://github.com/aatiibutt/TraffSign/.

5.1 Evaluation of Multi-lingual Text Detection Methods

The text detection dataset is split into 70% training, 20% test, and 10% validation sets. Initially, the EAST detector with the existing PA-net, VGG-16, and ResNet-50 as backbone architectures was evaluated. In the next step, the EAST detector with MobileNet as a backbone network is evaluated on the proposed dataset. The whole training process is executed for 100 epochs with the validation patience of 4. In addition, the Stochastic Gradient Descent (SGD) optimizer is set to 0.9, while the piece-wise learning rate and batch size are configured at 0.001 and 32, respectively. Moreover, the validation is performed after each epoch to monitor the learning progress. Performance has been in terms of precision, recall, and F-score of the evaluated networks.

Table 2. Comparison of text detection networks on proposed multilingual dataset demonstrating performance metrics. DB-Net and EAST with four backbone networks, PA-Net, VGG-16, ResNet-50, and improved MobileNet based architecture are evaluated after training with variable learning rates including piece-wise, 0.0001, and 0.001.

Model	Backbone	Learning rate	Precision	Recall	F-score
DB Net	FCN	Piece-wise	0.68	0.72	0.69
		0.0001	0.63	0.69	0.65
		0.001	0.61	0.66	0.62
EAST	PA-net	Piece-wise	0.51	0.39	0.44
		0.0001	0.54	0.42	0.46
		0.001	0.52	0.40	0.43
	VGG-16	Piece-wise	0.43	0.27	0.33
		0.0001	0.41	0.25	0.31
		0.001	0.39	0.24	0.29
	ResNet-50	Piece-wise	0.67	0.71	0.68
		0.0001	0.70	0.74	0.72
		0.001	0.64	0.68	0.69
	Mobile-Net	Piece-wise	**0.86**	**0.93**	**0.89**
		0.0001	0.82	0.89	0.85
		0.001	0.77	0.84	0.86

Discussion: It can be seen from Table 2 that DB-Net achieved 68% accuracy with piece-wise learning rate. Whereas, the EAST detector with PA-net which is base architecture, achieved 51% precision. However, it is worth noting that the same EAST detector with VGG-16 backbone architecture performed worst. Similarly, a base EAST detector with ResNet-50 architecture improved detection performance in terms of precision, recall, and F-score. While on the other side, our proposed improved EAST architecture with MobileNet outperformed the above-mentioned base architectures in terms of precision and recall i.e., 86% and

Fig. 7. Qualitative examples of proposed text detection network. The precise bounding boxes over Urdu and English text regions show that proposed network achieved significant performance in spotting multi-lingual text in test data.

93%, respectively. It is also important to mention that the learning rate played a crucial role in maximizing the performance of the networks. For instance, both architectures, i.e., DB and EAST performed well with piece-wise learning rates, while our presented MobileNet based EAST architecture achieved maximum precision. Some of the sample predictions of our presented network are depicted in Fig. 7.

5.2 Evaluation of Multi-lingual Text Recognition Methods

The proposed text recognition dataset is split into 70% training, 20% test, and 10% validation sets. CRNN-based architectures with LSTM and BLSTM layers comprising of variable hidden units have been employed while transfer learning is used to fine-tune these baseline models on our proposed dataset. The performance of the presented networks on our proposed dataset is measured in terms of Word Recognition Rate (WRR), as statistically elaborated in Table 3. It is worth mentioning that WRR is calculated by mapping the predicted characters over ground truth transcriptions and dividing the sum of correctly predicted characters with the sum of ground truth.

Discussion: It can be seen from Table 3 that baseline CRNN based text recognition methods achieved 71.22%, and 68.10%, WRR on English and Urdu scripts respectively. Moreover, baseline CRNN with BLSTM consisting of 256 hidden units achieved 78.34% and 76.41% WRR on English and Urdu scripts respectively. The proposed CRNN with BLSTM with 256 hidden units achieved 92.18%

Table 3. Performance of proposed CRNN Architectures on proposed multi-lingual text recognition dataset.

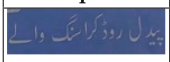
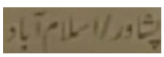
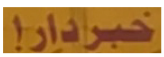
Architecture	Sequence learner	Language	Hidden units	WRR (%)
Baseline CRNN	LSTM	English	128	71.22
		Urdu		68.10
	BLSTM	English		74.87
		Urdu		72.29
	LSTM	English	256	75.41
		Urdu		72.66
	BLSTM	English		78.34
		Urdu		76.41
Improved CRNN	LSTM	English	128	82.64
		Urdu		79.28
	BLSTM	English		86.71
		Urdu		85.56
	LSTM	English	256	89.07
		Urdu		86.43
	BLSTM	English		92.18
		Urdu		90.85

and 90.85% WRR on English and Urdu respectively. Some of the sample predictions are shown in Table 4 to provide more insight into the influential factors. Among these, similar alphabets or special characters influenced the performance of proposed methods. For instance, referencing *output 79* in Table 4, our model is influenced by the English special character "**backward slash: /**", and predicted as Urdu alphabet " ا ". Moreover, in output 77 in Table 4, the proposed model predicts English special character **"exclamation mark: !"** as Urdu alphabet " ا " due to similarity. Another observation is that the proposed method demonstrated better applicability in recognizing text scripts of both languages; however, its performance is influenced by the presence of inter-ligature and intra-ligature overlapping. For example, in *output 21* in Table 4, our model ignored the Urdu alphabet " ت " in the word " رفتار " due to distant epigrams, written on next alphabet. Similarly, in *output 19*, proposed model missed " ن " in " پھینکا " and misrecognized " ع " in " منع " because of inter-ligature overlapping in multi-ligature based words. Whereas, it performed well with single ligature and two ligature-based words, as shown in *output 58, 39, 82*.

5.3 Evaluation of Proposed Text-Detection and Recognition Models as an End-to-End Pipeline

To evaluate the proposed methods as an end-to-end method, we combined both, the text detection and recognition models in a pipeline to evaluate in terms

Table 4. Qualitative examples of proposed CRNN architecture, evaluating on test set of proposed text recognition dataset.

Input	Recognized Text	WRR
	Out [74]: پیدل روڈ کراسنک والے	97.21%
	Out [21]: رفار آہستہ رکھیں	98.60%
	Out [19]: کوڑاکرکٹ پھینکنا منع ہے	88.57%
	Out [79]: پشاور ا اسلام آباد	98.12%
	Out [77]: خبردار ا	98.54%
	Out [33]: Hakla to DI Khan	100%
	Out [27]: PAF Academy Risalpur	100%

of inference time, taken in detecting and recognizing the text in test images. Overall, the proposed pipeline achieved 97.3% precision in text detection, 97.9% word recognition rate, with an average inference time of 0.17 s. Some of the sample predictions along with the recognized text and inference time are shown in Table 5. It can be observed from the table that the text detection network has demonstrated significant performance in spotting English and Urdu text regions. Precision in text detection is one of the most influential factors in the performance of text recognition networks. Subsequently, the text-recognition network also recognized the multi-lingual text with a minimal error rate with less inference time. However, the performance of text recognition is influenced to

Table 5. Performance of proposed text-detection and recognition models as an end-to-end pipeline (Matrices – Text-Detection: Precision)

Input Image	Text-Detection	Recognized Text	WRR (%)	Inf. Time (s)
	98.3%	اسلام آباد Islamabad Chakri چکری Lahore لاہور	100%	0.11 + 0.08
	98.7%	راولپنڈی Rawalpindi Peshawar پشاور Lahore لاہور اسلام آباد Islamabad	98.83%	0.09 + 0.06

some extent while recognizing Urdu alphabets with epigrams, written in overlapping and tilt orientation. These issues can be addressed by introducing text orientation-awareness in text recognition models.

6 Conclusions

In this paper, we present a large-scale multi-lingual dataset—DLL-TraffSiD for text detection and text recognition for Urdu/English traffic signboards. We also proposed a high-performance pipeline for multi-lingual text detection and text recognition tasks in scene images. The results show that the proposed text detection and recognition method outperformed the existing base architectures on our proposed dataset. As a future direction, we are aiming to extend this research work to text orientation assessment and context development.

References

1. Aberdam, A., et al.: Sequence-to-sequence contrastive learning for text recognition. In: CVPR, pp. 15302–15312 (2021)
2. Ali, A., Pickering, M., Shafi, K.: Urdu natural scene character recognition using convolutional neural networks. In: IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition, pp. 29–34. IEEE (2018)
3. Arafat, S.Y., Iqbal, M.J.: Urdu-text detection and recognition in natural scene images using deep learning. IEEE Access **8**, 96787–96803 (2020)
4. Bains, J.K., Singh, S., Sharma, A.: Dynamic features based stroke recognition system for signboard images of Gurmukhi text. Multimed. Tools Appl. **80**(1), 665–689 (2021)
5. Basavaraju, H.T., Manjunath Aradhya, V.N., Guru, D.S.: A novel arbitrary-oriented multilingual text detection in images/video. In: Satapathy, S.C., Tavares, J.M.R.S., Bhateja, V., Mohanty, J.R. (eds.) Information and Decision Sciences. AISC, vol. 701, pp. 519–529. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-7563-6_54
6. Bušta, M., Patel, Y., Matas, J.: E2E-MLT - an unconstrained end-to-end method for multi-language scene text. In: Carneiro, G., You, S. (eds.) ACCV 2018. LNCS, vol. 11367, pp. 127–143. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21074-8_11
7. Butt, M.A., Riaz, F.: CARL-D: a vision benchmark suite and large scale dataset for vehicle detection and scene segmentation. Signal Process.: Image Commun. **104**, 116667 (2022)
8. Chandio, A.A., Asikuzzaman, M., Pickering, M., Leghari, M.: Cursive-text: a comprehensive dataset for end-to-end Urdu text recognition in natural scene images. Data Brief **31**, 105749 (2020)
9. Chandio, A.A., Pickering, M.: Convolutional feature fusion for multi-language text detection in natural scene images. In: 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies, pp. 1–6. IEEE (2019)
10. Chandio, A.A., Pickering, M., Shafi, K.: Character classification and recognition for Urdu texts in natural scene images. In: 2018 International Conference on Computing, Mathematics and Engineering Technologies, pp. 1–6. IEEE (2018)

11. Chen, X., Jin, L., Zhu, Y., Luo, C., Wang, T.: Text recognition in the wild: a survey. ACM Comput. Surv. (CSUR) **54**(2), 1–35 (2021)
12. Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., Zhou, S.: Focusing attention: towards accurate text recognition in natural images. In: ICCV, pp. 5076–5084 (2017)
13. Ch'ng, C.K., Chan, C.S.: Total-text: a comprehensive dataset for scene text detection and recognition. In: 14th IAPR ICDAR, vol. 1, pp. 935–942. IEEE (2017)
14. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: CVPR, pp. 2315–2324 (2016)
15. Hossain, M.S., Alwan, A.F., Pervin, M.: Road sign text detection using contrast intensify maximally stable extremal regions. In: 2018 IEEE Symposium on Computer Applications & Industrial Electronics, pp. 321–325. IEEE (2018)
16. Karatzas, D., et al.: ICDAR 2015 competition on robust reading. In: 13th ICDAR, pp. 1156–1160. IEEE (2015)
17. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: AAAI, vol. 34, pp. 11474–11481 (2020)
18. Long, S., He, X., Yao, C.: Scene text detection and recognition: the deep learning era. Int. J. Comput. Vision **129**(1), 161–184 (2021)
19. Lu, N., et al.: MASTER: multi-aspect non-local network for scene text recognition. Pattern Recogn. **117**, 107980 (2021)
20. Nayef, N., et al.: ICDAR 2019 robust reading challenge on multi-lingual scene text detection and recognition—RRC-MLT-2019. In: 2019 ICDAR, pp. 1582–1587. IEEE (2019)
21. Panhwar, M.A., Memon, K.A., Abro, A., Zhongliang, D., Khuhro, S.A., Memon, S.: Signboard detection and text recognition using artificial neural networks. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication, pp. 16–19. IEEE (2019)
22. Rasib, M., Butt, M.A., Riaz, F., Sulaiman, A., Akram, M.: Pixel level segmentation based drivable road region detection and steering angle estimation method for autonomous driving on unstructured roads. IEEE Access **9**, 167855–167867 (2021)
23. Reddy, S., Mathew, M., Gomez, L., Rusinol, M., Karatzas, D., Jawahar, C.: RoadText-1K: text detection & recognition dataset for driving videos. In: 2020 ICRA, pp. 11074–11080. IEEE (2020)
24. Saha, S., et al.: Multi-lingual scene text detection and language identification. Pattern Recogn. Lett. **138**, 16–22 (2020)
25. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell. **39**(11), 2298–2304 (2016)
26. Shi, B., et al.: ICDAR 2017 competition on reading Chinese text in the wild (RCTW-17). In: 14th IAPR ICDAR, vol. 1, pp. 1429–1434. IEEE (2017)
27. Sun, Y., Liu, J., Liu, W., Han, J., Ding, E., Liu, J.: Chinese street view text: large-scale Chinese text reading with partially supervised learning. In: ICCV, pp. 9086–9095 (2019)
28. Tian, S., et al.: Multilingual scene character recognition with co-occurrence of histogram of oriented gradients. Pattern Recogn. **51**, 125–134 (2016)
29. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: COCO-text: dataset and benchmark for text detection and recognition in natural images. arXiv (2016)
30. Yuan, T.L., Zhu, Z., Xu, K., Li, C.J., Mu, T.J., Hu, S.M.: A large Chinese text dataset in the wild. J. CS Technol. **34**(3), 509–521 (2019)
31. Zhou, X., et al.: EAST: an efficient and accurate scene text detector. In: CVPR, pp. 5551–5560 (2017)