IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Towards AI-Enabled Approach for Urdu Text Recognition: A Legacy for Urdu Image Apprehension

KAMLESH NARWANI[1,2], HONGZHI LIN [3], SANDEEP PIRBHULAL[4], AND MIR HASSAN.[5,6],

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology(HUST), 1037 Luoyu Rd, Hongshan, Wuhan, 430074, Hubei, China (e-mail: i201522249@hust.edu.cn)
[2]Faculty of Science and Technology, ILMA University, Korangi Creek, Karachi, 75190, Sindh, Pakistan
[3]Faculty of Electronic Information and Communications Department, Huazhong University of Science andTechnology (HUST), 1037 Luoyu Rd, Hongshan, Wuhan, 430074, Hubei, China (e-mail: linhongzhi1@hotmail.com)
[4]Dept. of Information Security and Communication Technology Norwegian University of Science andTechnology (NTNU) Critical Infrastructure Security and Resilience group, NO-7491, Høgskoleringen, 7491, Trondheim, Norway (e-mail: s.pirbhulal@hotmail.com)
[5]Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China (e-mail: hassanmir@hust.edu.cn)
[6]Institure of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania

Corresponding author: Kamlesh Narwani (e-mail: i201522249@hust.edu.cn).

ABSTRACT Recognizing Urdu text in natural images is more challenging as compared to other languages, such as English, due to the cursive nature of Urdu script. However, Urdu scene text has not received enough attention from both industry and academia due to the lack of the dataset of Urdu text. We propose a large-scale Urdu Scene Text Dataset (USTD) to address this problem, which is designed for Urdu scene text detection and recognition. The proposed dataset contains 29674 text annotations (17877 Urdu and 11797 English), 749725 characters in 6389 images. It covers a wide variety of text images with both Nastaleeq and Naskh writing styles, taken from different streets and roads of Pakistan. The vast diversity of this dataset makes it a benchmark to work on and train robust neural networks for the detection and recognition of cursive text. Besides, baseline results are also provided with several state-of-the-art networks, including TextBoxes++, Seglink, DB(ResNet-50) and EAST for text localization and Convolutional Recurrent Neural Network (CRNN) for text recognition. To further evaluate the performance of these models, we have used the most popular evaluation matrices of precision, recall, and F-measure. Our experimental outputs reveal that an end-to-end combination of DB(ResNet-50) and CRNN provides the best results with precision, recall, and F-measure of 0.7526, 0.5974, and 0.6660, respectively.

INDEX TERMS Cursive text recognition, Deep networks, End-to-end networks, Scene text dataset, Text localization, Urdu scene text.

I. INTRODUCTION

Text extraction from natural images has been the center of attention for the research community in computer vision. It has the potential to be used in a variety of real-world applications, such as assisting visually impaired persons [1], autonomous traffic sign recognition [2], scene understanding [3], robot navigation [4] and license plate detection [5]. Researchers have well-studied text extraction in documents, and many commercial products are available, having recognition accuracy of more than 99% [6] on the documented text. However, the complex structure of scene images makes the identification of text a very challenging task. Unlike scanned documents, scene images come with different challenges such as sensor noise, blur, varying resolution, non-planer objects, unknown layouts, inconsistent lighting conditions, arbitrary angled and occluded or distorted text, as presented in Fig 1.

The lack of dataset availability adds to the difficulties of text extraction in the natural world, which work as an essential ingredient to train robust deep networks. Latin based and some other languages, such as English, Chinesh, French, and Korean, have received enough attention from the research community. Several datasets such as ICDAR Robust Reading [7], SVHN [8], COCO-Text [9], FSNS [10], KAIST [11], RCTW-17 [12] and CTW [13] have proved to be an essential factor for improved performance on these languages.

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3203426

(a) Blurred (b) Complex background (c) Non-planer

(d) Occluded (e) Inconsistent light

FIGURE 1: Challenges of natural scene text

Whereas, the cursive text is yet to be analyzed thoroughly due to the unavailability of any standard dataset. Urdu is one of the major languages of South Asia. It is Pakistan's national language and is also spoken in most parts of India. With over 170 million speakers, it is one of the most common language used in the world. Urdu script is cursive in nature, and recognizing cursive text is more challenging and difficult than extracting Latin text due to the complex structure of its characters and its ligature based nature, where characters are combined to make words and sentences. Unlike English, Urdu script is written in opposite direction, from right to left, and it has two main styles: Nastaleeq and Naskh. Nastaleeq is a Perso-Arabic script which is a flowy and ornate and hanging script. Whereas, Naskh is a slightly angular and stodgy script that comes from Arabic. Magazines, newspapers, and books mostly follow Nastaleeq writing style, whereas most of the online content, such as the content on bbcurdu.com, follow Naskh writing style. The diagonal and overlapping nature of Nastaleeq makes it occupy less space for a ligature. A visual comparison of both writing styles is given in Fig 2. Unlike Naskh writing style, characters in Nastaleeq mostly overlap each other, making character segmentation unsuitable.

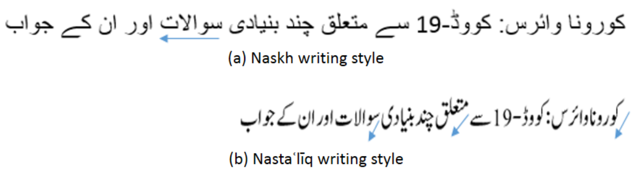


FIGURE 2: Visual comparison of Naskh and Nastaleeq writing styles

Another complication in Urdu writing is that the shape of the character varies depending on its position (start, middle or end) of the ligature. Fig 3 shows four different characters in blue, red, green and light blue colors that change their shapes depending on their position of occurrence. A character may have as many as 60 different shapes [14], so the exact classification and extraction of these characters is a strenuous task.

FIGURE 3: Different shapes of same characters based on their position in ligature

In this work, Urdu Scene Text Dataset (USTD), a large dataset of cursive text in scene images is presented. It is, to our knowledge, the largest dataset on Urdu text in imagery. It contains 6389 images with 749725 Urdu and English characters and 29674 text instances. Unlike many publicly available datasets, where images are taken with the assistance of Google Street View [8], [10], [15] or Tanscent Street View [13], most images in USTD are taken with a mobile camera at different streets and roads of Pakistan. It contains text in both the writing styles of Nastaleeq and Naskh, which makes it a diverse and complex dataset. Annotations are provided for both Urdu and English content in each image. For each text, we annotate its content, its bounding box coordinates, and an attribute to represent if its Urdu or English text.
Several state-of-the-art deep networks are trained on this dataset for text detection and recognition. Since these are

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3203426

the first models to be used on this dataset so these networks provide baseline results. It is expected that the depth, diversity, and complexity of this dataset will make it the most suitable dataset to train deep networks for Urdu text detection and recognition. Towards this end, we have summarized our contribution in three folds:

1) We first prepared the largest scene text dataset of the Urdu script for algorithm development and comparison. The benchmark contains 6389 scene images with 29674 text annotations of Urdu and English text. Images come from diverse real-world situations such as signboards, billboards, and shop names.
2) To assess the challenges of proposed benchmark, four regression and segmentation based cutting-edge text detection approaches are used and baseline results are provided for both Urdu and English text, separately.
3) Several experiments are conducted and end-to-end recognition results are provided using four different models, which prove the usefulness of USTD. The experimental results show that the proposed dataset has a promising aspect for any future work to be done on Urdu scene text recognition.

The remainder of this paper is structured in the following manner. Related work is discussed in Section 2, while Section 3 delves into the specifics of our dataset. Section 4 discusses the state-of-the-art algorithms used the proposed dataset, as well as their output matrices. The experimental findings and discussion are presented in Section 5. Finally, Section 6 provides a summary of the entire work.

II. RELATED WORK

Disregarding robust deep networks and computing devices, the computer vision community is still struggling at the extraction of text in natural images. The lack of publicly accessible datasets to manipulate is a major explanation for this underachievement. This paper presents the world's largest dataset of both cursive and Latin based (Urdu and English) text in natural scene images, as well as baseline findings from a number of cutting-edge techniques. Therefore, this discussion is confined to similar datasets and methods for extracting text from natural images. Discussion is restricted to text in natural images because the extraction of documented text has already been well studied.

A. DATASETS OF TEXT IN NATURAL IMAGE

A proper sized and well-annotated dataset plays a vital role to exploit any computer vision algorithm or classifier to its fullest. ICDAR 2003 [7] was the first competition to create the basis in the field of image text detection and recognition. It comprises 509 scene text images with most of the text content appearing at the center of images. Later on, a series of scene text datasets [16]–[18] were released by ICDAR, with each possessing different challenges. ICDAR 2015 [19] includes the most difficult images, referred to as incidental text, which were all captured using Google glasses with little regard for image quality. During that period, many other datasets [20]–[22] were also made publicly available and have proved to be standard benchmarks to evaluate the performance of computer vision algorithms. Table 1 summarizes the stats of few of the most popular and non-cursive scene text datasets publicly available.It includes datasets with horizontal text (HT), arbitrary quadrilateral text (AQT), irregular text (IT), and synthetic text (Syn).

Despite several publicly available datasets, most of the work focuses on English text or numbers. Urdu or Urdu like scripts, such as Persian and Arabic, have attracted the least attention in this field. Few attempts have been made to capture and prepare datasets for cursive text in scene images and videos. Authors in [34] worked on recognizing multilingual text in natural scenes, capturing 1100 Urdu scene text images and combining them with data from ICDAR 2017-MLT [30]. In [35], [36] authors have worked on character classification and recognition of Urdu text, but the number of images used is under 850, and segmented characters are below 18000. Urdu news ticker detection and recognition have been worked out in [37], [38], where authors have collected video images from different channels in both high and low quality. However, the text in news ticker images generally appears at either bottom or top on images, which makes the text localization task easier. The largest Urdu scene text dataset is presented in [39], where author has collected 2500 natural outdoor images with three different languages text, Urdu, English and Sindhi. This dataset is further processed to get cropped isolated characters and word dataset. Considering the tedious task of training deep neural network to recognize scene text, this dataset doesn't seem to be enough. Synthetic Urdu text is presented in [40], where author has generated 51K synthetic images with embedded Urdu text. It contains 1600 unique ligatures with each ligature having 32 variations. Apparently it seems to be a huge dataset to train a deep network but synthetic data can not take place of text in the wild. This dataset maybe used to pre-train a model to further improve the performance.

Arabic Text has also found some interest from the research community. ARASTI [41] and ARASTEC [42] both present Arabic text datasets, but the size of both datasets makes them unsuitable for benchmarking. In [43]–[45], Arabic text samples are gathered from various news channels, e.g., BBC Arabic, France 24 Arabic, Al Jazeera, and Al Arabiya. Despite all these attempts, it appears that more efforts are required to make a standard benchmark dataset as far as the detection and recognition of Urdu scene text is concerned. Realizing its potential value, the first publicly available largest dataset on Urdu scene text is proposed. Table 2 compares USTD with other cursive text datasets so far used.

B. TEXT DETECTION AND RECOGNITION

Reading text in the wild can be divided into two sub-tasks: text detection and text recognition. First, the presence of text is detected by localizing its position in character/word bounding boxes followed by text recognition in which the localized/cropped text is transcribed into a machine-readable

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3203426

TABLE 1: Details of few most popular non-cursive scene text benchmark datasets. EN and CN stand for English and Chinese languages, respectively. Whereas, Train, Valid and Test represent training, validation and testing sets

Dataset	Year	Number of Images			Script Type	Layout
		Train	Valid	Test		
IC03 [7]	2003	258	-	251	EN Text	HT
SVT [23]	2010	100	-	249	EN Text	AQT
SVHN [8]	2010	73257	-	26032	EN Digits	HT
IC11 [17]	2011	229	-	233	EN Text	HT
MSRA-TD500 [20]	2012	300	-	200	EN, CN Text	AQT
III-T5K-Word [21]	2012	2000	-	3000	EN Text	AQT
IC13 [18]	2013	229	-	233	EN Text	HT
USTB-SV1K [24]	2013	1000(Total)	-	-	EN Text	AQT
SVTP [25]	2013	-	-	639	EN Text	AQT
CUTE [26]	2014	-	-	80	EN Text	IT
IC15 [19]	2015	1000	-	500	EN Text	AQT
SynthText [27]	2016	800k(Total)	-	-	EN Text	Syn
COCO-Text [9]	2017	43686	10000	10000	EN Text	AQT
CTW [13]	2017	25000	-	6000	CN Text	AQT
RCTW-17 [12]	2017	11514	-	1000	CN Text	AQT
ToT [28]	2017	1255	-	300	CN, EN Text	IT
SCUT-CTW1500 [29]	2017	1000	-	500	EN Text	IT
MLT17 [30]	2017	7200	1800	9000	9 Languages	AQT
ArTs19 [31]	2019	5603	-	4563	CN, EN Text	IT
MLT19 [32]	2019	10000	-	10000	10 Languages	AQT
LSVT19 [33]	2019	20157	4968	4841	CN Text	IT

TABLE 2: Comparison of cursive text datasets so far used

Name / Author	**Content**	**Data Size**	**Size**	**Script**	**Availability**
Chandio *et al.* [39]	Scene Text	2500 Images	13778 words	Urdu, English and Sindhi	Available
Arafat *et al.* [40]	Synthetic Text	51k Images	51k words	Urdu	Available
Chandio *et al.* [34]	Scene Text	1000/100 Images	-/-	Urdu and English	Unavailable
Ali *et al.* [46]	Scene Text	845 Images	28000 Segmented characters	Urdu and English	Unavailable
Sami-Ur-Rehman *et al.* [37]	Video Frames	News tickers from 41 channels	20097 tickers	Urdu	Unavailable
Raza *et al.* [38]	Video Images	1000 Images	23833 words	Urdu and English	Available
Ahmed *et al.* [47]	Scene Text	2469 Images	19300 characters, and 7765 words	Arabic and English	Unavailable
Tounsi *et al.* [41]	Scene Text	374 Images	2093 characters	Arabic and English	Available
Urdu Scene Text Dataset (USTD)	**Scene Text**	**6389 Images**	**29674 text lines and 749725 characters**	**Urdu and English**	**Will be made available**

form. This whole problem is addressed in three different manners by the research community as text detection, text recognition, and an integrated approach known as end-to-end recognition

1) Text Detection

The process of identifying the presence of text using character/word bounding boxes is known as text detection. Before the incorporation of deep convolutional neural networks, traditional text detection approaches required scheming and testing a vast number of likely handcrafted features. The traditional approaches were usually based on either Stroke width transform (SWT) [48], [49], or maximally stable extremal regions (MSERs) [50]–[53]. SWT is an image operator that takes an image and outputs a new equally sized stroke-width image, where every single element relates to the pixel value of each stroke width. One of the best features of SWT is that it is language-independent, i.e., it can detect the script of any language, but its limitation is that it is best suited for only clean text. MSER takes an image and extracts its MSER regions in the original image, whereas non-textual candidates are disposed of by using filters. Contrary to traditional approaches, deep convolutional neural networks based approaches [54], [55] enjoy the luxury of automatically detecting features which overall simplifies the pipeline of text detection [56], [57].

These approaches are further classified into two classes: 1) Regression-based method. Object detection models like SSD [58], which directly regress the bounding box of the targeted object, influenced these methods [59]–[61]. Unlike general objects, text appear with different orientation, shapes and non-uniform aspect ratios. Due to which, object detection frameworks cannot be directly used for text detection. TextBoxes++ [62] results quadrilateral regression of the text by adjusting anchor ratios and changing convolutional kernels in SSD. An attention based technique is proposed in SSTD [63] to roughly spot text areas. DeepReg [64] and EAST [56] target multi-directional text by resulting pixel-level regression. Regression-based methods are not burdened with heavy post-processing algorithms and can efficiently detect text with varying aspect ratios at higher inference speed. These approaches, however, frequently fail to detect multi-oriented text, such as curved text. 2) Segmentation-based method. These approaches are mostly based on semantic segmentation methods, and they get the bounding box of text by cascading pixel-level prediction information and using a post-processing algorithm. Number of these methods [65]–[69] employ fully convolutional network to extract the segmented text area. These methods can efficiently detect arbitrary shaped text but usually suffer slower inference speed due to heavy dependence on post-processing algorithms. Apart from that their performance also banks on the quality of segmentation accuracy.

2) Text Recognition

Once the text is detected in imagery, text recognition transcribes the localized text into the machine-interpretable form. It can further be categorized into character-based [70], [71] and word-based recognition [72], [73]. Unlike English, the cursive script is a ligature based script where characters are combined to make words and sentences. Additionally, in the Nastaleeq writing style, words are written in a diagonal manner where most of the characters overlap with each other, making character-based recognition an unpopular approach for Urdu text.

3) End-to-end Recognition

End-to-end recognition systems [74], [75], sometimes termed as an integrated approach, takes an image with a complex background, localizes and detects the presence of any text instance, and finally converts imagery text into human understandable strings. End-to-end recognition guarantees satisfactory performance, which is usually compromised due to error propagation between detection and recognition in two-step methods [76].

III. PROPOSED DATASET DESCRIPTION

In this section we present a vast dataset on Urdu scene text, namely, Urdu Scene Text Dataset (USTD). We will discuss the data acquisition procedure and the proposed ground truth annotation approach. The organization of dataset and statistical analysis, are also presented here.

A. DATA ACQUISITION OF USTD

USTD is composed of 6389 Urdu scene text images. Since English is the official language of Pakistan, so its quite normal to find English script in most of the images. To maintain maximum content diversity, we have collected images occurring in various scenarios like signboards, billboards, street and shop names, advertisement banners, etc. Few samples of USTD are presented in Fig 4.

Around 70% of images are captured by a mobile camera from different cities in Pakistan, whereas the remaining 30% are taken from Google images, Facebook, and other sources. Since images are not collected from a uniform source, so they come with varying image quality, which is later pre-processed as per network requirements. As images are captured by individuals, so there might be a possibility of having duplicated areas. Initially, we collected 8450 images, which were later reduced to 6389 on the criteria of having more than 70% duplicated area.

B. GROUND TRUTH ANNOTATIONS

The proper declaration of ground truths plays a vital role in supervised learning techniques. Improper and misleading ground truths can lead to vague results. Generally, three different approaches are used when annotating ground truths of text images, namely, character level, word level, and line-level annotations. Ligature based nature of Urdu makes character-level recognition very difficult because most of the characters overlap each other as depicted in Fig 5(a). Additionally, Nastaleeq script usually occurs with an uneven length of white space between intra-ligature and inter-ligatures (see Fig 5 (b)), making word-level recognition also challenging.

Ground truths in USTD are manually annotated in line-level granularity with a quadrilateral shape. We have used simple annotation tool with the Urdu keyboard to translate the ground truth text in UTF-8 encoded .txt file. The annotated text file contains coordinates of each text line bounding boxes and its transcription, as presented in Fig 6.

USTD contains both Urdu and English text, so we have used an attribute (0 for Urdu and 1 for English) for each script for modularity convenience. In case, if one prefers to evaluate the recognition of Urdu text only, one could use this attribute to filter out instances with English text and vice versa. Numbers that appear separately are considered as English text; otherwise, if it occurs in the middle or along with Urdu script, then it is considered as Urdu text.

C. DATASET ORGANIZATION AND STATISTICS

The proposed dataset has been split into two sections: training set and testing set. 80% of the total images (5100 out of 6389) are allocated for training the models, while the other 1289 photos are devoted to the testing set. In order to avoid any similarity, we have manually checked both sets, and any instance of having similar images has been removed. After splitting, the training set contains 23724 text instances (14549 Urdu text lines and 9175 English text lines), and the testing set comprises 5950 text instances (3328 Urdu text lines and 2622 English text lines). A detailed distribution of the number of text occurrences in images can be viewed in Fig 7. While most of the images contain below 15 text occurrences, but few images have more than 50 text lines. On average, each image comprises 4.6 text lines.

FIGURE 4: Few samples of USTD images

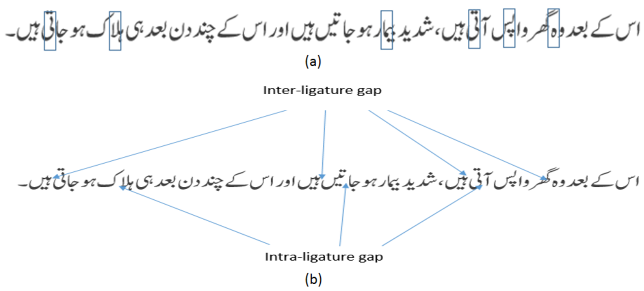


FIGURE 5: Complexities of Nastaleeq writing style

FIGURE 6: Images in USTD and their corresponding ground truth annotations

IV. BASELINE ALGORITHMS AND EVALUATION PROTOCOLS

This section presents the baseline algorithms used of text localization and recognition. It also discusses evaluation protocols used to measure the performance of these algorithms on USTD.

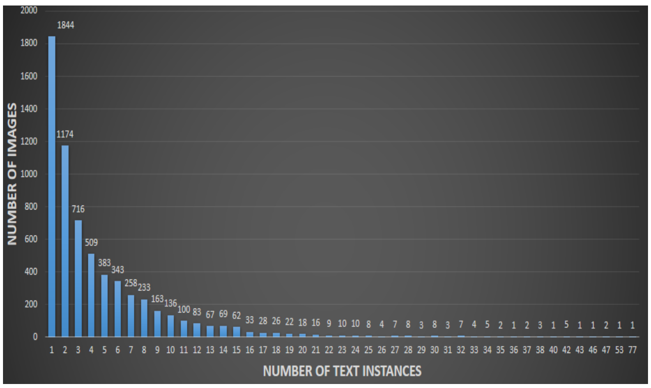


FIGURE 7: Text distribution in USTD dataset

A. BASELINE ALGORITHMS

Text detection entails detecting or localizing any text instances in a given image, usually in the shape of a bounding box. For the task of text detection, we have adopted four most popular state-of-the-art networks, namely, TextBoxes++ [62], EAST [56], SegLink [77], and Differentiable Binarization (DB) [78].

TextBoxes++ is a fully trainable end-to-end regression-based text detector. It is inspired by SSD [58], an object detection model. But unlike SSD, which uses rectangular box representation for detecting objects, TextBoxes++ employees quadrilateral or oriented rectangle representation for scene text. Apart from that, TextBoxes++ utilizes long convolutional kernels to get better receptive field for predicting the bounding boxes. It is highly efficient at spotting non-horizontal scene text. Unlike conventional text detectors, which comprise multiple stages such as, generation, filtering and grouping of character or word candidates, TextBoxes++ is setup on a simpler pipeline. With just minor modifications, the VGG-16 [79] architecture serves as its backbone. Overall, it's a fully-convolutional network with only convolutional and pooling layers for predicting the bounding boxes of text instances. To generate final outputs, the predicted bounding boxes are passed into non-maximum suppression (NMS). The only post-processing stage is NMS.

EAST is a regression-based, simple and powerful text detector, which is efficient at detecting the text with arbitrary orientation without requiring any intermediate stage. It is based on DenseBox [80] and consists of only two stages, an FCN and an NMS. The pixel-level score map of several channels and geometry is generated using FCN. FCN produces text areas, which are subsequently supplied to NMS to generate final results.

SegLink is also a regression-based and fully-convolutional neural network based model. It manoeuvre single-shot multi-box detector SSD [58] to detect text instances with the help of segments and links. The total count of oriented bounding boxes on a text instance is measured in segments, and links are utilized to cascade segments that belong to the same word or line. SegLink uses VGG-16 as its backbone model.

Differentiable Binarization (DB) is a module used along with segmentation-based network to efficiently detect scene text with arbitrary angles (including curved text). Traditionally segmentation based networks rely on post-processing algorithms to obtain bounding boxes from probability maps generated by segmentation networks. DB instead uses the concept of joint optimization by amalgamating the binarization process with segmentation network. It utilizes differentiable binarization (an approximate function of binarization) to optimize segmentation network in training process. DB employees ResNet-50 [81] as the backbone of segmentation network. It also utilizes modulated deformable convolution, as used in [82] , for flexible receptive field for the network. Modulated deformable convolutions are used at multiple convolutional layers of ResNet-50.

The driving force behind selecting these networks is that they are robust at detecting oriented text. Unlike conventional text, text in imagery can occur in multi-orientation. While many other algorithms struggle at detecting non-horizontal text, TextBoxes++, EAST, SegLink, and DB perform better at spotting uneven text, hence reducing any ambiguity in results (see [56], [62], [77], [78]for more details).

Given the bounding box, text recognition transcribes the bounded text into a machine and human-understandable form. In this work, we have used CRNN [72], an state-of-the-art text recognition method. It is an end-to-end trainable neural network that takes an image as input and results in the recognized text. It is consists of three different layers, convolutional layer, recurrent layer, and transcription layer for feature extraction, label distribution prediction, and frame prediction, respectively (see [72] for detail).

B. EVALUATION MATRICES

We used the most generally employed precision, recall, and F-measure matrices to evaluate the performance of state-of-the-art baseline algorithms for text recognition. Precision is the proportion of correctly detected occurrences among all instances that should have been detected, whereas recall can be interpreted as the proportion of correctly identified occurrences among all instances that should have been identified. F-measure combines precision and recall to represent the system's total accuracy. Low false positives and false negatives are indicators of a successful F-measure. Precision P, recall R, and F-measure F are all calculated using the following formulas,

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

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F = 2 \times \frac{P \times R}{P + R} \tag{3}$$

The number of hit detection boxes, mismatched detections, and undiscovered text boxes are given by TP, FP, and FN, respectively. TP returns 0 or 1 depending on the overlap threshold chosen between the detected box and ground truth box, and it is given by the following equation,

If $\frac{(A_{GT} \cap A_{DT})}{(A_{GT} \cup A_{DT})}$>TH then TP = 1, else TP = 0.

where A_{GT} and A_{DT} are ground truth and detection areas, threshold TH is also sometimes termed as intersection-over-union IoU, which is normally set to 0.5.

V. EXPERIMENTAL RESULTS AND DISCUSSION

Implementation details: Instead of pre-training the detection models on any other dataset, we train them with our training dataset. For DB, poly learning rate is used and it varies with each iteration. For an iteration, the learning rate (lr_current) is set by is set by following formula.

$$lr\_current = lr\_initial \times (1 - \frac{iter}{max\_iter})^{power} \tag{4}$$

Where lr_initial is set to 0.007 and power is 0.9. The parameters of momentum and Weight decay are set to 0.9 and 0.0001, respectively. EAST is trained with Adam [83] optimizer. VGG is employed as the backbone model and BatchNorm2d is used for normalization. Instead of using balanced cross entropy, dice loss is used to optimize IoU of segmentation. SGD algorithm is used to optimize the Seglink and momentum is set to 0.9. TextBoxes++ is also trained

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2022.3203426

with Adam [83] and it is implemented in two stages. In first stage the model is trained with our training data at learning rate 10^{-4} and then in second stage training is continued at smaller learning rate and higher negative ratio. For CRNN, ADADELTA [84] is used for optimization purpose, which automatically sets learning rate. Table **??** summarizes the implementation details of all four text detectors.

Experimental environment: TextBoxes++ is implemented using Caffe, EAST and DB are implemented using PyTorch, and Tensor Flow is employed for Seglink. Whereas, for recognition model Torch is used. All the experimentation are carried on a workstation with a 3.5 GHz Intel(R) Xeon(R) CPU E5-2637 v4, 64 GB RAM, and TITAN Xp.

A. RESULTS

Initially, the performance of all four text detectors is evaluated on our proposed USTD dataset. Few qualitative based results are presented in Fig 8, where all the correct and false positive results are represented in red bounding boxes and non detected text regions are not bounded by any box.

Following the stated evaluation protocols, quantitative results are given in Table 4, under IoU threshold levels of 0.5 and 0.7. The performance of all four models reduces considerably with a higher threshold levels of IoU because higher IoU requires accurate and robust text detectors. In comparison to the other three models, DB performs better.

TABLE 4: Detection results of Seglink [77], EAST [56], TextBoxes++ [62] and Differential Binarization [78]

Method	IoU Threshold = 0.5			IoU Threshold = 0.7		
	P	R	F	P	R	F
Seglink	0.6904	0.5064	0.5843	0.506	0.461	0.482
EAST	0.5915	0.5142	0.5501	0.534	0.473	0.501
TextBoxes++	0.6134	0.6926	0.6505	0.582	0.496	0.535
DB	0.8364	0.6920	0.7574	0.7541	0.6208	0.6809

USTD contains both Urdu and English text, with 60% of text instances being Urdu and the remaining 40% are English text. So we have evaluated the performance of all four detectors separately on Urdu and English text (IoU is set to 0.5) and summarized it in Table 5.

TABLE 5: Detection results of Seglink [77], EAST [56], TextBoxes++ [62] and Differential Binarization [78] on Urdu and Egnlish text separately

Method	Urdu Text			English Text		
	P	R	F	P	R	F
Seglink	0.588	0.476	0.526	0.826	0.721	0.769
EAST	0.592	0.478	0.529	0.829	0.748	0.786
TextBoxes++	0.598	0.484	0.535	0.836	0.753	0.792
DB	0.7179	0.6147	0.6623	0.8701	0.7412	0.8005

After analyzing the detection results of all four models, it is observed that these models can detect English text quite efficiently but struggle at spotting Urdu text because Urdu text has higher aspect ratio as compared to English text. Apart from that Urdu characters are very similar to some symbols, which may lead to missed detection or error detection. Urdu

(a) SegLink results

(b) EAST results

(c) TextBoxes++ results

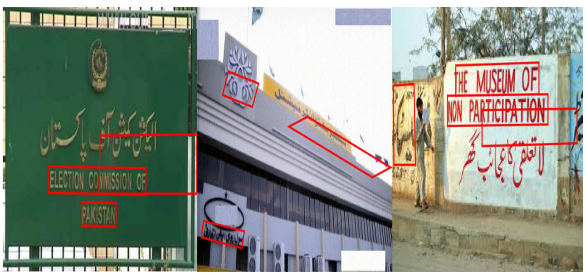
(d) DB results

FIGURE 8: Qualitative comparison of text detection performance on some USTD images. Detection results (correct and false positive) are presented by red bounding boxes, whereas missed detection are not bounding by any box

script is also easily effected by the background. Few false detection (false positive and false negative) results are presented in Fig 9.

(a) SegLink results

(b) EAST results



(c) TextBoxes++ results

(d) DB results

FIGURE 9: Few examples of false positive and false negative results.

The above statement can further by validated by Table 6, where we have compared the performance of stated models on two other datasets (containing only English Text), i-e, ICDAR 2015 Incidental Text [19] and COCO Text [9]. From obtained results it can be witnessed that like COCO Text, USTD also presents various difficulties because of its diversity, and each of the four detectors performs poorly on USTD as compared to the Incidental Text dataset of ICDAR 2015 (IC15). A significant reason for the worst performance on COCO Text is that the dataset was collected without considering text in mind [9].

Text detection without correct recognition is meaningless because it only fulfills half of the goal of scene text reading. To accomplish the task of text recognition, we have used the CRNN model [72] separately with each text detector in an end-to-end recognition manner . Some of the quantitative results for each end-to-end recognition model are shown in Fig 10.

(a) SegLink results

(b) EAST results

(c) TextBoxes++ results

(d) DB results

FIGURE 10: Qualitative comparison of end-to-end text recognition on some USTD images. Green bounding boxes represent detected text and recognized text is given in yellow color

As witnessed in Table 7 , the overall performance of all end-to-end recognition models reduce by some margin because CRNN is not trained with any lexicon. It can be observed that end-to-end recognition model of DB and CRNN outperforms all other models by at least 14% (F-measure).

Based on these findings, it can be concluded that, despite employing various state-of-the-art text localization and recognition approaches that perform rather well on other publicly accessible datasets, none of them are successful in recognizing Urdu text in USTD. Compared to the regression-

TABLE 6: Text localization results on ICDAR 2015 Incidental Text (IC15), COCO Text, and USTD. NR stands for Not Reported

Method	Dataset								
	IC15			COCO Text			USTD		
	P	R	F	P	R	F	P	R	F
Seglink	0.768	0.731	0.750	0.312	0.501	0.384	0.6904	0.5064	0.5843
EAST	0.783	0.833	0.807	0.324	0.504	0.394	0.5915	0.5142	0.5501
TextBoxes++	0.785	0.878	0.829	0.567	0.608	0.587	0.6134	0.6926	0.6505
DB	0.918	0.832	0.873	NR	NR	NR	0.8364	0.6920	0.7574

TABLE 7: End-to-end recognition results on USTD

Method	Task	Performance		
		P	R	F
SegLink	Detection	0.6904	0.5064	0.5843
	Detection + Recognition	0.5916	0.4341	0.5007
EAST	Detection	0.5915	0.5142	0.5501
	Detection + Recognition	0.5513	0.4096	0.4700
TextBoxes++	Detection	0.6134	0.6926	0.6505
	Detection + Recognition	0.5678	0.4909	0.5265
DB	Detection	0.8364	0.6920	0.7574
	Detection + Recognition	0.7526	0.5974	0.6660

based models, segmentation-based text detector performs better.

VI. CONCLUSION AND FUTURE WORK

We presented for the first time a large dataset of Urdu scene text images (to be made publicly available) with the goal of improving Urdu natural scene text recognition. It includes 6389 images, carrying total text content of 29674 Urdu and English text lines. We have annotated each image, where an annotated text file contains the coordinates of text instances, transcribed content, and its attribute, indicating whether it is Urdu or English text. Baseline results are also presented for various state-of-the-art text identification methods. From our outcomes, it tends to be dissected that, however, these models exceed expectations at recognizing English text, yet every one battle at perceiving Urdu content. We intend to construct an end-to-end recognition network in the future, focusing on the complexities of the Urdu script. We also plan to organize the very first competition for the recognition of Urdu scene text where the different contestants can submit, assess, and equate their work. We are confident that future work in Urdu text detection and recognition will be greatly influenced by this dataset.

• • •

REFERENCES

[1] X. Zhang, X. Liu, T. Sarkodie-Gyan, and Z. Li, “Development of a character CAPTCHA recognition system for the visually impaired community using deep learning,” Machine Vision and Applications, vol. 32, no. 1, pp. 1–19, 2021, company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Springer Berlin Heidelberg. [Online]. Available: https://sci-hub.se/https://link.springer.com/article/10.1007/s00138-020-01160-8

[2] S.-K. Tai, C. Dewi, R.-C. Chen, Y.-T. Liu, X. Jiang, and H. Yu, “Deep Learning for Traffic Sign Recognition Based on Spatial Pyramid Pooling with Scale Analysis,” Applied Sciences, vol. 10, no. 19, p. 6997, 2020, number: 19 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2076-3417/10/19/6997

[3] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso, and A. Torralba, “Semantic Understanding of Scenes Through the ADE20K Dataset,” International Journal of Computer Vision, vol. 127, no. 3, pp. 302–321, 2019.

[4] Y. Chen, C. Liu, B. E. Shi, and M. Liu, “Robot Navigation in Crowds by Graph Convolutional Networks With Attention Learned From Human Gaze,” IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 2754–2761, 2020, conference Name: IEEE Robotics and Automation Letters.

[5] W. Weihong and T. Jiaoyang, “Research on License Plate Recognition Algorithms Based on Deep Learning in Complex Environment,” IEEE Access, vol. 8, pp. 91 661–91 675, 2020, conference Name: IEEE Access.

[6] J. J. Weinman, E. Learned-Miller, and A. R. Hanson, “Scene Text Recognition Using Similarity and a Lexicon with Sparse Belief Propagation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 10, pp. 1733–1746, 2009.

[7] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, “ICDAR 2003 robust reading competitions,” in Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., vol. 1. Edinburgh, UK: IEEE Comput. Soc, 2003, pp. 682–687. [Online]. Available: http://ieeexplore.ieee.org/document/1227749/

[8] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning,” 2011.

[9] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, “COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images,” arXiv:1601.07140 [cs], 2016, arXiv: 1601.07140. [Online]. Available: http://arxiv.org/abs/1601.07140

[10] R. Smith, C. Gu, D.-S. Lee, H. Hu, R. Unnikrishnan, J. Ibarz, S. Arnoud, and S. Lin, “End-to-End Interpretation of the French Street Name Signs Dataset,” in Computer Vision – ECCV 2016 Workshops, ser. Lecture Notes in Computer Science, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, pp. 411–426.

[11] J. Jung, S. Lee, M. S. Cho, and J. H. Kim, “Touch TT: Scene Text Extractor Using Touchscreen Interface,” ETRI Journal, vol. 33, no. 1, pp. 78–88, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.4218/etrij.11.1510.0029

[12] B. Shi, C. Yao, M. Liao, M. Yang, P. Xu, L. Cui, S. Belongie, S. Lu, and X. Bai, “ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17),” in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, 2017, pp. 1429–1434, iSSN: 2379-2140.

[13] T.-L. Yuan, Z. Zhu, K. Xu, C.-J. Li, and S.-M. Hu, “Chinese Text in the Wild,” arXiv:1803.00085 [cs], 2018, arXiv: 1803.00085. [Online]. Available: http://arxiv.org/abs/1803.00085

[14] S. Hussain, “Complexity of asian writing systems: a case study of nafees nasta’leeq for urdu,” in Proceedings of the 12th AMIC Annual Conference on e-Worlds: Governments, Business and Civil Society, Asian Media Information Center, Singapore. Citeseer, 2003.

[15] K. Wang, B. Babenko, and S. Belongie, “End-to-end scene text recognition,” in 2011 International Conference on Computer Vision, 2011, pp. 1457–1464, iSSN: 1550-5499.

[16] S. Lucas, “ICDAR 2005 text locating competition results,” in Eighth International Conference on Document Analysis and Recognition (ICDAR’05), 2005, pp. 80–84 Vol. 1, iSSN: 2379-2140.

[17] D. Karatzas, S. R. Mestre, J. Mas, F. Nourbakhsh, and P. P. Roy, “ICDAR 2011 Robust Reading Competition - Challenge 1: Reading Text in Born-Digital Images (Web and Email),” in 2011 International Conference on Document Analysis and Recognition, 2011, pp. 1485–1490, iSSN: 2379-2140.

[18] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i. Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazàn, and L. P. de las Heras,

"ICDAR 2013 Robust Reading Competition," in 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 1484–1493, iSSN: 2379-2140.

[19] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on Robust Reading," in 2015 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, pp. 1156–1160.

[20] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1083–1090, iSSN: 1063-6919.

[21] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in BMVC-British Machine Vision Conference. BMVA, 2012.

[22] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition," arXiv:1406.2227 [cs], 2014, arXiv: 1406.2227. [Online]. Available: http://arxiv.org/abs/1406.2227

[23] K. Wang and S. Belongie, "Word Spotting in the Wild," in Computer Vision – ECCV 2010, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer, 2010, pp. 591–604.

[24] X. Yin, X. Yin, K. Huang, and H. Hao, "Robust Text Detection in Natural Scene Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 5, pp. 970–983, 2014, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[25] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 569–576.

[26] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," Expert Systems with Applications, vol. 41, no. 18, pp. 8027–8048, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414004060

[27] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2315–2324.

[28] C. K. Ch'ng and C. S. Chan, "Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, 2017, pp. 935–942, iSSN: 2379-2140.

[29] Y. Liu, L. Jin, S. Zhang, C. Luo, and S. Zhang, "Curved scene text detection via transverse and longitudinal sequence connection," Pattern Recognition, vol. 90, pp. 337–345, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320319300664

[30] N. Nayef, F. Yin, I. Bizid, H. Choi, Y. Feng, D. Karatzas, Z. Luo, U. Pal, C. Rigaud, J. Chazalon, W. Khlif, M. M. Luqman, J.-C. Burie, C.-l. Liu, and J.-M. Ogier, "ICDAR2017 Robust Reading Challenge on Multi-Lingual Scene Text Detection and Script Identification - RRC-MLT," in 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Kyoto: IEEE, 2017, pp. 1454–1459. [Online]. Available: http://ieeexplore.ieee.org/document/8270168/

[31] C. K. Chng, Y. Liu, Y. Sun, C. C. Ng, C. Luo, Z. Ni, C. Fang, S. Zhang, J. Han, E. Ding, J. Liu, D. Karatzas, C. S. Chan, and L. Jin, "ICDAR2019 Robust Reading Challenge on Arbitrary-Shaped Text - RRC-ArT," in 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1571–1576, iSSN: 2379-2140.

[32] N. Nayef, Y. Patel, M. Busta, P. N. Chowdhury, D. Karatzas, W. Khlif, J. Matas, U. Pal, J. Burie, C. Liu, and J. Ogier, "ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition — RRC-MLT-2019," in 2019 International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1582–1587, iSSN: 2379-2140.

[33] Y. Sun, J. Liu, W. Liu, J. Han, E. Ding, and J. Liu, "Chinese street view text: Large-scale chinese text reading with partially supervised learning," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9086–9095.

[34] A. A. Chandio and M. Pickering, "Convolutional Feature Fusion for Multi-Language Text Detection in Natural Scene Images," in 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). Sukkur, Pakistan: IEEE, 2019, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8673517/

[35] A. A. Chandio, M. Pickering, and K. Shafi, "Character classification and recognition for Urdu texts in natural scene images," in 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET). Sukkur: IEEE, 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8346341/

[36] M. A. Panhwar, K. A. Memon, A. Abro, D. Zhongliang, S. A. Khuhro, and S. Memon, "Signboard Detection and Text Recognition Using Artificial Neural Networks," in 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC). Beijing, China: IEEE, 2019, pp. 16–19. [Online]. Available: https://ieeexplore.ieee.org/document/8784625/

[37] Sami-Ur-Rehman, B. U. Tayyab, M. F. Naeem, A. Ul-Hasan, and F. Shafait, "A Multi-faceted OCR Framework for Artificial Urdu News Ticker Text Recognition," in 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). Vienna: IEEE, 2018, pp. 211–216. [Online]. Available: https://ieeexplore.ieee.org/document/8395197/

[38] A. Raza and I. Siddiqi, "A database of artificial urdu text in video images with semi-automatic text line labeling scheme," in Proc. 4th Int. Conf. Adv. Multimedia (MMEDIA), 2012, pp. 75–81.

[39] A. A. Chandio, M. Asikuzzaman, M. Pickering, and M. Leghari, "Cursive-Text: A Comprehensive Dataset for End-to-End Urdu Text Recognition in Natural Scene Images," Data in Brief, vol. 31, p. 105749, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920306430

[40] S. Y. Arafat and M. J. Iqbal, "Urdu-Text Detection and Recognition in Natural Scene Images Using Deep Learning," IEEE Access, vol. 8, pp. 96 787–96 803, 2020, conference Name: IEEE Access.

[41] M. Tounsi, I. Moalla, and A. M. Alimi, "ARASTI: A database for Arabic scene text recognition," in 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). Nancy, France: IEEE, 2017, pp. 140–144. [Online]. Available: http://ieeexplore.ieee.org/document/8067776/

[42] M. Tounsi, I. Moalla, A. M. Alimi, and F. Lebouregois, "Arabic characters recognition in natural scenes using sparse coding for feature representations," in 2015 13th International Conference on Document Analysis and Recognition (ICDAR). Tunis, Tunisia: IEEE, 2015, pp. 1036–1040. [Online]. Available: http://ieeexplore.ieee.org/document/7333919/

[43] S. Yousfi, S.-A. Berrani, and C. Garcia, "ALIF: A dataset for Arabic embedded text recognition in TV broadcast," in 2015 13th International Conference on Document Analysis and Recognition (ICDAR). Tunis, Tunisia: IEEE, 2015, pp. 1221–1225. [Online]. Available: http://ieeexplore.ieee.org/document/7333958/

[44] O. Zayene, J. Hennebert, S. Masmoudi Touj, R. Ingold, and N. Essoukri Ben Amara, "A dataset for Arabic text detection, tracking and recognition in news videos- AcTiV," in 2015 13th International Conference on Document Analysis and Recognition (ICDAR). Tunis, Tunisia: IEEE, 2015, pp. 996–1000. [Online]. Available: http://ieeexplore.ieee.org/document/7333911/

[45] M. Jain, M. Mathew, and C. V. Jawahar, "Unconstrained scene text and video text recognition for Arabic script," in 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). Nancy, France: IEEE, 2017, pp. 26–30. [Online]. Available: http://ieeexplore.ieee.org/document/8067754/

[46] A. Ali, M. Pickering, and K. Shafi, "Urdu Natural Scene Character Recognition using Convolutional Neural Networks," in 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), 2018, pp. 29–34.

[47] S. B. Ahmed, S. Naz, M. I. Razzak, and R. B. Yusof, "A Novel Dataset for English-Arabic Scene Text Recognition (EASTR)-42K and Its Evaluation Using Invariant Feature Extraction on Detected Extremal Regions," IEEE Access, vol. 7, pp. 19 801–19 820, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8641268/

[48] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 2963–2970, iSSN: 1063-6919.

[49] A. Mosleh, N. Bouguila, and A. B. Hamza, "Image Text Detection Using a Bandlet-Based Edge Detector and Stroke Width Transform," in Procedings of the British Machine Vision Conference 2012. Surrey: British Machine Vision Association, 2012, pp. 63.1–63.12. [Online]. Available: http://www.bmva.org/bmvc/2012/BMVC/paper063/index.html

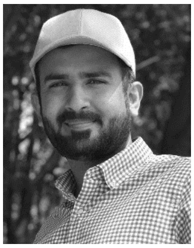
[50] M. S. Extremal, J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust Wide Baseline Stereo from," in In British Machine Vision Conference, 2002, pp. 384–393.

[51] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions," in 2011 18th IEEE International Conference on Image Processing, 2011, pp. 2609–2612, iSSN: 1522-4880.
[52] H. I. Koo and D. H. Kim, "Scene Text Detection via Connected Component Clustering and Nontext Filtering," IEEE Transactions on Image Processing, vol. 22, no. 6, pp. 2296–2305, 2013.
[53] L. Neumann and J. Matas, "A Method for Text Localization and Recognition in Real-World Images," in Computer Vision – ACCV 2010, ser. Lecture Notes in Computer Science, R. Kimmel, R. Klette, and A. Sugimoto, Eds. Berlin, Heidelberg: Springer, 2011, pp. 770–783.
[54] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4159–4167.
[55] Y. Liu, L. Jin, and C. Fang, "Arbitrarily Shaped Scene Text Detection With a Mask Tightness Text Detector," IEEE Transactions on Image Processing, vol. 29, pp. 2918–2930, 2020, conference Name: IEEE Transactions on Image Processing.
[56] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: an efficient and accurate scene text detector," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 5551–5560.
[57] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, "TextBoxes: A Fast Text Detector with a Single Deep Neural Network," in Thirty-First AAAI Conference on Artificial Intelligence, 2017. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14202
[58] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in Computer Vision – ECCV 2016, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.
[59] S. Wang, Y. Liu, Z. He, Y. Wang, and Z. Tang, "A quadrilateral scene text detector with two-stage network architecture," Pattern Recognition, vol. 102, p. 107230, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320320300364
[60] L. Deng, Y. Gong, Y. Lin, J. Shuai, X. Tu, Y. Zhang, Z. Ma, and M. Xie, "Detecting multi-oriented text with corner-based region proposals," Neurocomputing, vol. 334, pp. 134–142, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231219300256
[61] Y. Cai, W. Wang, H. Ren, and K. Lu, "SPN: short path network for scene text detection," Neural Comput & Applic, vol. 32, no. 10, pp. 6075–6087, 2020. [Online]. Available: https://doi.org/10.1007/s00521-019-04093-0
[62] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A Single-Shot Oriented Scene Text Detector," IEEE Transactions on Image Processing, vol. 27, no. 8, pp. 3676–3690, 2018, conference Name: IEEE Transactions on Image Processing.
[63] P. He, W. Huang, T. He, Q. Zhu, Y. Qiao, and X. Li, "Single shot text detector with regional attention," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3047–3055.
[64] W. He, X.-Y. Zhang, F. Yin, and C.-L. Liu, "Deep direct regression for multi-oriented scene text detection," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 745–753.
[65] D. Deng, H. Liu, X. Li, and D. Cai, "PixelLink: Detecting Scene Text via Instance Segmentation," AAAI, vol. 32, no. 1, 2018, number: 1. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/12269
[66] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, and S. Shao, "Shape robust text detection with progressive scale expansion network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9336–9345.
[67] Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, and J. Jia, "Learning shape-aware embedding for scene text detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4234–4243.
[68] Y. Xu, Y. Wang, W. Zhou, Y. Wang, Z. Yang, and X. Bai, "TextField: Learning a Deep Direction Field for Irregular Scene Text Detection," IEEE Transactions on Image Processing, vol. 28, no. 11, pp. 5566–5579, 2019, conference Name: IEEE Transactions on Image Processing.
[69] S. Long, J. Ruan, W. Zhang, X. He, W. Wu, and C. Yao, "Textsnake: A flexible representation for detecting text of arbitrary shapes," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 20–36.
[70] C. Shi, C. Wang, B. Xiao, Y. Zhang, S. Gao, and Z. Zhang, "Scene Text Recognition Using Part-Based Tree-Structured Character Detection," 2013, pp. 2961–2968. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2013/html/Shi_Scene_Text_Recognition_2013_CVPR_paper.html
[71] C. Yao, X. Bai, B. Shi, and W. Liu, "Strokelets: A Learned Multi-Scale Representation for Scene Text Recognition," 2014, pp. 4042–4049. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2014/html/Yao_Strokelets_A_Learned_2014_CVPR_paper.html
[72] B. Shi, X. Bai, and C. Yao, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298–2304, Nov. 2017.
[73] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading Text in the Wild with Convolutional Neural Networks," Int J Comput Vis, vol. 116, no. 1, pp. 1–20, Jan. 2016. [Online]. Available: https://doi.org/10.1007/s11263-015-0823-z
[74] J. J. Weinman, Z. Butler, D. Knoll, and J. Feild, "Toward Integrated Scene Text Reading," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 2, pp. 375–387, Feb. 2014.
[75] L. Neumann and J. Matas, "Scene Text Localization and Recognition with Oriented Stroke Detection," 2013, pp. 97–104. [Online]. Available: http://openaccess.thecvf.com/content_iccv_2013/html/Neumann_Scene_Text_Localization_2013_ICCV_paper.html
[76] S. Long, X. He, and C. Yao, "Scene Text Detection and Recognition: The Deep Learning Era," Int J Comput Vis, vol. 129, no. 1, pp. 161–184, Jan. 2021. [Online]. Available: https://doi.org/10.1007/s11263-020-01369-0
[77] B. Shi, X. Bai, and S. J. Belongie, "Detecting Oriented Text in Natural Images by Linking Segments," in 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. IEEE Computer Society, 2017, pp. 3482–3490. [Online]. Available: https://doi.org/10.1109/CVPR.2017.371
[78] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-Time Scene Text Detection with Differentiable Binarization," AAAI, vol. 34, no. 07, pp. 11 474–11 481, 2020, number: 07. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6812
[79] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556 [cs], 2015, arXiv: 1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556
[80] L. Huang, Y. Yang, Y. Deng, and Y. Yu, "DenseBox: Unifying Landmark Localization with End to End Object Detection," arXiv:1509.04874 [cs], 2015, arXiv: 1509.04874. [Online]. Available: http://arxiv.org/abs/1509.04874
[81] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[82] X. Zhu, H. Hu, S. Lin, and J. Dai, "Deformable convnets v2: More deformable, better results," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9308–9316.
[83] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv:1412.6980 [cs], 2017, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980
[84] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," arXiv:1212.5701 [cs], Dec. 2012, arXiv: 1212.5701. [Online]. Available: http://arxiv.org/abs/1212.5701

KAMLESH NARWANI received his M.E degree from the school of Information and Telecommunication Engineering, University of Science and Technology Beijing, China. He is currently a Ph.D. student with the school of Electronic Information and Communication, HUST. He is also working as a Lecturer at the Faculty of Science and Technology, ILMA University, Karachi, Pakistan. His research interests include scene text detection and recognition.

HONGZHI LIN received the BS, MS, and Ph.D. degrees from the Huazhong University of Science and Technology, China, in 2000, 2003, and 2008, respectively. He is currently an assistant professor with the Huazhong University of Science and Technology. His current research interests include the areas of wireless networking and digital image processing.

SANDEEP PIRBHULAL received his Ph.D. degree in Pattern Recognition and Intelligent Systems from the University of Chinese Academy of Sciences in 2017. He is currently working as a postdoctoral researcher at the Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Norway. Dr. Pirbhulal has a vast experience of 7 years in Academia and Research. His current research focus on critical infrastructure security, network security, information security, privacy and security for WSNs, wireless body sensor networks (WSNs), and Internet of Things. He has published more than 60 international journal articles including IEEE TBME, IEEE Sensor Journal, and IEEE JBHI among others, over 15 international conference proceeding papers, and 4 book chapters.



MIR HASSAN is pursuing a PhD in Computer Architecture at Wuhan National laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China. He did master of engineering in software engineering from Wuhan University, Wuhan, China. Currently, he is serving as Junior researcher at Vilnius University, Vilnius, Lithuania. He also served as Visiting Postgarduate Researcher at School of Computing Science, University of Glasgow, UK. His research interest are: Blockchain Technology, Internet of Things, and Machine learning.