Indic Scene Text on the Roadside

Ajoy Mondal[0000−0002−4808−8860], Krishna Tulsyan[0000−0003−0177−588X], and C V Jawahar[0000−0001−6767−7057]

CVIT, International Institute of Information Technology, Hyderabad, India
krishna.tulsyan@research.iiit.ac.in
{ajoy.mondal,jawahar}@iiit.ac.in

Abstract. Extensive research and the development of benchmark datasets have primarily focused on Scene Text Recognition (STR) in Latin languages. However, the scenario differs for Indian languages, where the complexities in syntax and semantics have posed many challenges, resulting in limited datasets and comparatively less research in this domain. Overcoming these challenges is crucial for advancing scene text recognition in Indian languages. Although a few works have touched upon this issue, they are constrained in the size and scale of the data as far as we know. To bridge this gap, this paper introduces a large scale, diverse dataset, named as *IIIT-IndicSTR-Word* for Indic scene text. Comprising a total of 250K word level images in ten different languages — *Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, and Telugu*, these images are extracted from roadside scenes captured by a GoPro camera. The dataset encompasses a wide array of realistic adversarial conditions, including blur, changes in illumination, occlusion, non-iconic texts, low resolution, and perspective text. We establish a baseline for the proposed dataset, facilitating evaluation and benchmarking with a specific focus on STR tasks. Our findings indicate that our dataset is a practical training source to enhance performance on respective datasets. The code, dataset, and benchmark results are available at https://cvit.iiit.ac.in/usodi/istr.php.

Keywords: Scene Text Recognition (STR), word images, Indic languages, Indic scene text, roadside, benchmark.

1 Introduction

Language serves as a universal medium for global communication, facilitating exchanges and interactions among people worldwide. The diverse array of languages across different communities highlights the recognition of language's crucial role in human connectivity. As a written form of language, text significantly enhances the potential for information transfer. In the wild, where writing manifests semantic richness, valuable information that holds the key to understanding the contemporary environment is embedded. The textual information found in diverse settings is pivotal in various applications, ranging from image search and translation to transliteration, assistive technologies (especially for the visually impaired), and autonomous navigation. In the modern era, the automatic extraction of text from photographs or frames depicting natural environments, known as Scene Text Recognition (STR) or Photo-OCR, is a significant challenge. This complex problem is typically divided into two sub-problems: scene text detection, which involves locating text within an image, and cropped word image recognition.

Optical Character Recognition (OCR) has traditionally concentrated on interpreting printed or handwritten text within documents. However, the proliferation of capturing devices like mobile

Fig. 1. Showcases valuable sources of scene text in various Indic languages within roadside images captured by a camera.

Dataset	Word Images	Language	Features	#Language	#IpL
MLT-19 [17]	191K	Multi-lingual [1]	Irr	10	19.1K
MLT-17 [18]	96K	Multi-lingual[2]	Irr	9	10.6K
Urdu-Text [3]	14K	Urdu	Irr, Noisy, LR	1	14K
IndicSTR12 [14]	27K	Multi-lingual [3]	Irr, LR, Blur, Occ, PT	12	2.25K
IIIT-IndicSTR-Word	250K	Multi-lingual [4]	Irr, LR, Blur, Occ, PT	10	25K

Table 1. Statistics regarding several publicly available real scene text recognition datasets are presented. #Language denotes the total number of languages in each dataset, while #IpL represents the average number of word level images per language. Irr, LR, Occ, and PT indicate the presence of irregular text, low resolution images, occlusion, and perspective text, respectively.

phones and video cameras has underscored the significance of scene text recognition (STR), presenting a challenge whose resolution holds substantial potential for advancing various applications.

[1] Arabic, Bengali, Chinese, Devanagari, English, French, German, Italian, Japanese, and Korean
[2] Arabic, Bengali, Chinese, English, French, German, Italian, Japanese, and Korean
[3] Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu, and Urdu
[4] Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu

Despite the heightened interest in STR within the research community, it comes with unique challenges, including varying backgrounds in natural scenes, diverse scripts, fonts, layouts, styles, and image imperfections related to text, such as blurriness, occlusion, and uneven illumination.

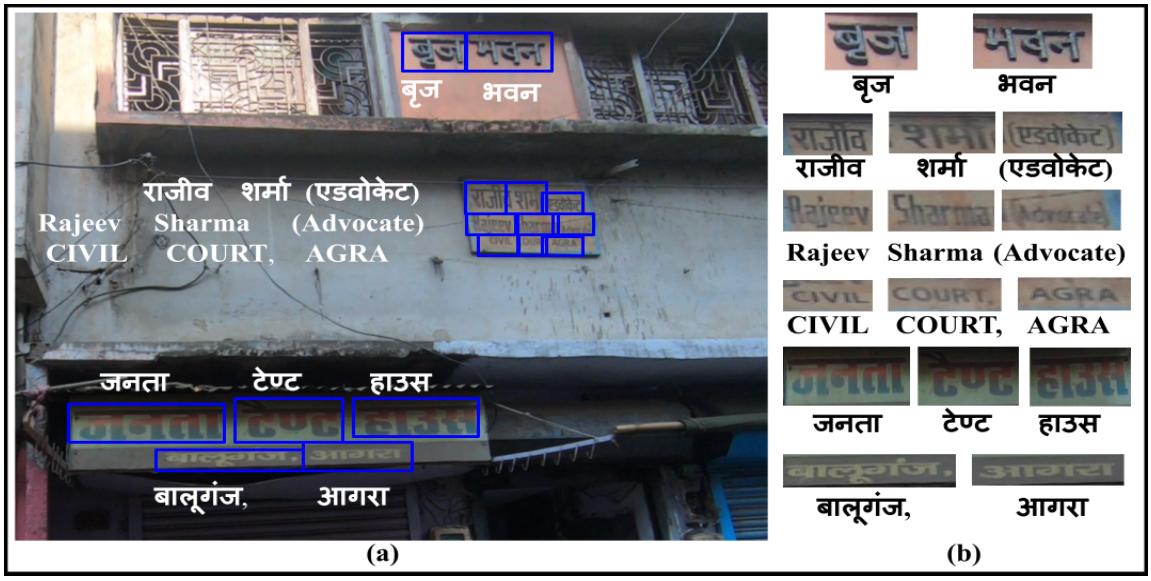


Fig. 2. (a) Depicts a single annotated image frame from our dataset. (b) Represents cropped word level images and their corresponding transcriptions serve as ground truths.

Fig. 3. Illustrates a few sample word level images of each of the ten languages from our dataset.

In response to these challenges, researchers have taken on the task of curating datasets tailored to address specific issues, each highlighting unique features and representing subsets of challenges

encountered in real-world scenarios. As part of this effort, several benchmark datasets focused on Latin scripts have been developed, including IIIT5K-Words [16], SVT [29], ICDAR2003 [13], ICDAR2013 [8], ICDAR2015 [7], SVT Perspective [19], and CUTE80 [21]. Additionally, there are multi-lingual datasets such as MTWI [5], LSVT [25,26], and MLT-17 [18], which serve as valuable resources for creating and evaluating scene text recognition (STR) models.

In the context of Indic languages, relatively few efforts have focused on datasets and models. For instance, MLT-19 [17] (featuring only Bengali and Devanagari), MLT-17 [18] (limited to Bengali), Urdu-Text [3], and IndicSTR12 [14] (covering twelve Indic languages) are some examples. IndicSTR12 is the largest Indic scene text recognition dataset, encompassing twelve languages spoken in India. However, it contains a limited number of word level images (approximately 2K-3K) per language, which needs to be increased for training deep STR models. Consequently, there is a growing need for a more extensive and diverse dataset explicitly designed for STR in Indic languages. This demand has emerged to meet the evolving research requirements in scene text recognition for Indic scripts.

To bridge this gap, our paper introduces a large scale, diverse dataset, named as *IIIT-IndicSTR-Word* for scene text recognition in Indic languages. Roadside scene images are a rich resource of scene text images for STR tasks. Illustrated in Fig. 1 are Indian roadside scene images captured across different states. These images exhibit diverse text characteristics, including occlusion or partial occlusion, variations in illumination, motion blur, perspective distortion, orientation differences, and variations in style, size, color, and language. They serve as invaluable assets for constructing expansive Indic scene text recognition datasets. Leveraging this resource, we meticulously capture numerous roadside scene images from diverse Indian states using a GoPro camera. Subsequently, we extract words from these images and annotate them with corresponding transcriptions, which serve as ground truths [5] (refer Fig. 2). The resulting dataset encompasses 250K word level images spanning ten languages: *Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu*. Additionally, we present a high-performing baseline for STR in Indic languages. In summary, our contributions can be outlined as follows:

- We present *IIIT-IndicSTR-Word*, a vast and diverse Indic scene text recognition dataset. It comprises 250K word level images extracted from roadside scenes captured by GoPro cameras. Encompassing ten major languages of India — *Bengali, Gujarati, Hindi, Kannada, Malayalam, Oriya, Punjabi, Tamil, and Telugu* (refer to Fig. 3), our dataset provides comprehensive coverage. For a thorough overview and comparison with existing datasets, please consult Table 1. The statistics demonstrate that our dataset offers remarkable diversity and is notably larger than existing Indic STR datasets. To our knowledge, it is the most extensive dataset for Indic scene text recognition.
- The images in our dataset, encompass a wide range of variations, including differences in font styles, low resolution, partial occlusion, perspective imaging, illumination variations, varying text lengths, and multi-orientation across languages (refer Fig. 4). This diverse collection is pivotal in developing robust and high-performing Indic scene text recognition models.
- We offer a baseline model for the scene text recognition task. The results showcase how the dataset enhances the performance of the model (refer Table 4 and Table 5). Our findings sug-

[5] The focus of this work is primarily on datasets and baselines for word level models. However, this effort will help in the future with many other tasks, including (i) script or language separation, (ii) text detection, (iii) end-to-end recognition, etc. One needs script separation, language classification, and scene text detection annotations for end-to-end recognition of these images, which are not part of this work. However, it will be publicly available.

gest that our dataset is a valuable training resource to enhance performance on corresponding datasets.

2 Related Work

Scene text recognition (STR) models commonly rely on Convolutional Neural Networks (CNNs) to encode image features. Decoding text from these learned image features typically involves two main approaches: Connectionist Temporal Classification (CTC) or an encoder-decoder framework with an attention mechanism. CTC-based methods, as seen in works such as [24,11], treat images as sequences of vertical frames and combine per-frame predictions using specific rules to generate the complete text. In contrast, encoder-decoder frameworks, like the one presented in [12], use attention to align input and output sequences. Both CTC and attention-based models have been explored extensively in STR research. DTRN [6] is noteworthy for being one of the first to employ CRNN models, a fusion of CNNs with stacked RNNs, to generate convolutional feature slices for RNN feeding. Using attention [12] introduces an STR-based encoder-decoder model, where the encoder is trained with binary constraints to reduce computational overhead.

Indic Scene Text Recognition: The shortage of annotated data poses a significant hurdle, especially in Indian languages, making it challenging to replicate the success seen in Latin STR solutions. Over the years, various efforts have been made to tackle this issue, though they have been sporadic and language-specific in Indian languages. Notably, the work by [3] marks the first endeavor to introduce a dataset and assess STR performance tailored explicitly for Urdu text. The MLT-17 dataset [18] encompasses 18k scene images across multiple languages, including Bengali. Its successor, MLT-19 [17], extends this to 20k scene images covering Bengali and Hindi, thus becoming the sole multi-lingual dataset encompassing ten languages. Additionally, Minesh *et al.* [15] train a CRNN model on synthetic data for Malayalam, Hindi, and Telugu, introducing the IIIT-ILST dataset for testing these languages. Meanwhile, Patel *et al.* [2] propose a CNN and CTC-based approach for script identification, text localization, and recognition for Bengali using the MLT-17 dataset. An OCR-on-the-go model [22] achieves a WRR of 51.01% on the IIIT-ILST Hindi dataset and a CRR of 35% on a multi-lingual dataset comprising 1000 videos in English, Hindi, and Marathi. Furthermore, Gunna *et al.*[4] delve into transfer learning among Indian languages to enhance WRR, introducing a dataset of natural scene images in Gujarati and Tamil to test this hypothesis, resulting in WRR improvements on the IIIT-ILST dataset for Gujarati and Tamil, respectively. More recently, Lunia *et al.* [14] introduced the IndicSTR12 dataset and established a high-performing baseline for STR across twelve Indic languages.

3 IIIT-IndicSTR-Word Dataset

We gather several thousand diverse roadside scene images using GoPro cameras from two to three metropolitan cities across various states of India. Primarily we cover states — *West Bengal, Gujarat, Delhi, Karnataka, Kerala, Maharashtra, Odisha, Punjab, Tamil Nadu, and Telangana* to capture scene text images of corresponding languages — *Bengali, Gujarati, Hindi, Kannada, Malayalam, Oriya, Punjabi, Tamil, and Telugu*. We capture images during day times. Subsequently, we extract word images from these scene images and manually generate transcripts for these word images as ground truth data for the text recognition task. This process leads to the creation of data corpus,

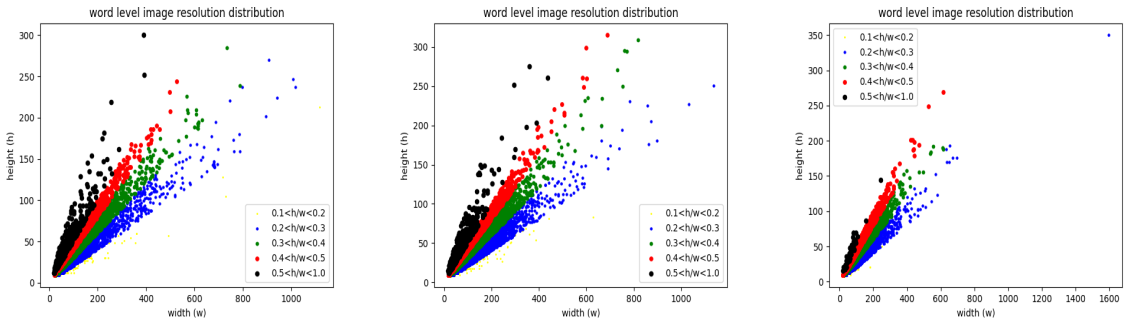
Fig. 4. The images showcased in this display exhibit various word samples sourced from our dataset. These words are extracted from roadside scene images captured by a GoPro camera. The collection encompasses a variety of word images, showcasing features such as partial occlusion, low resolution, font variation, perspective text, illumination variation, and multi oriented texts.

comprising 250K word level images representing ten popular Indic languages: *Bengali, Gujarati, Hindi, Kannada, Malayalam, Oriya, Punjabi, Tamil, and Telugu.* To represent the ground truth data corpus, we utilize the standard XML format, which includes the names of word level images and their corresponding textual transcriptions. Fig. 2 presents a sample annotation for reference. In Fig. 2(a), depicts a single annotated image frame from our dataset, while Fig. 2 represents cropped word level images, and their corresponding transcriptions serve as ground truths.

3.1 Dataset Feature and Statistics

Diversity: As scene text images in ten distinct languages are captured from various states of India, the resulting dataset is remarkably diverse. Fig. 4 provides a glimpse of this diversity through a selection of sample word images spanning multiple languages from our dataset. These images encompass a wide range of variations, including differences in font styles, low resolution, partial occlusion, perspective imaging, illumination variations, varying text lengths, and multi orientation across languages. This diverse collection is pivotal in developing robust and high-performing Indic

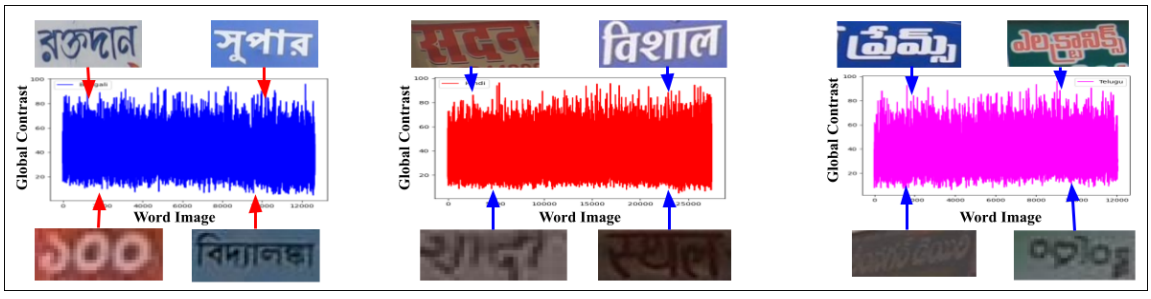
scene text recognition models. By encompassing such a broad spectrum of linguistic and visual characteristics, the dataset serves as a valuable resource for advancing research in this field and enhancing the accuracy and versatility of scene text recognition systems.



(a) Bengali (b) Hindi (c) Telugu

Fig. 5. Shows word level image resolution distribution for three languages Bengali, Hindi and Telugu.

Word Level Image Resolution Distribution: The variation in language, font, and geographical regions across India contributes to the diversity in the resolution of scene text words. This diversity in word resolution significantly improves the model's ability to generalize. Fig. 5 illustrates the distribution of resolutions in word level images, showcasing the dataset's extensive variability. It is evident from the figure that the resolution of word level images varies across different languages[6]. Incorporating words with diverse resolutions enriches the dataset, allowing the model to accommodate a broader range of visual characteristics and enhancing its performance across various styles and imaging conditions.



(a) Bengali (b) Hindi (c) Telugu

Fig. 6. Shows varying global contrast among word images in our dataset for three languages Bengali, Hindi and Telugu.

[6] Additional plots are available in the supplementary material

Contrast of Word Images: As word images are extracted from roadside scene images captured by GoPro cameras, the intensities vary significantly from one word image to another, resulting in fluctuations in contrast. To assess the ease of recognition for each word image, we employ the global contrast strategy [10]. Fig. 6 showcases the diverse global contrast levels among word images in Bengali, Hindi, and Telugu languages, respectively[7]. The figures demonstrate that contrast levels vary between 20 and 70 across the three languages. This variability in intensity within word images increases the complexity of the dataset and contributes to the creation of robust STR models. In contrast, understanding and leveraging these variations can enhance the adaptability and effectiveness of STR algorithms across diverse linguistic contexts. Additionally, exploring similar analyses for other languages included in the dataset can offer further insights into its characteristics and implications for STR model performance.

Language	**#Train**	**#Val**	**#Test**	**#Total**	**Word Length**		
					Min	**Max**	**Avg**
Bengali	17500	2500	5000	25000	1	14	4
Gujarati	17500	2500	5000	25000	1	15	4
Hindi	17500	2500	5000	25000	1	19	6
Kannada	17500	2500	5000	25000	1	19	6
Malayalam	17500	2500	5000	25000	1	19	6
Marathi	17500	2500	5000	25000	1	19	6
Oriya	17500	2500	5000	25000	1	19	5
Punjabi	17500	2500	5000	25000	1	19	5
Tamil	17500	2500	5000	25000	1	19	5
Telugu	17500	2500	5000	25000	1	19	7
Total	175000	25000	50000	250000	-	-	-

Table 2. The breakdown of our dataset into training, validation, and testing sets is presented for each language.

Dataset Split: Due to limitation of real Indic STR datasets, synthetic scene text [14] plays an important role for pre-training deep architecture. However, limited real scene text images are not able to reduce the domain gap between real and synthetic scene text images, resulting in poor recognition accuracy in real situation. To keep in mind, we divide our dataset containing 250K word images into 175K word images for training, 25K word images for validation, and 50K word images for testing. Table 2 shows statistics of our dataset. It will ensure that the models are trained on a large and diverse dataset and will help to improve their accuracy and performance.

3.2 Comparison with Existing Datasets

Table 3 compares our dataset and existing Latin, multi-lingual, and Indic scene text recognition datasets, highlighting significant differences and advantages. Various factors, such as dataset size and diversity in word level images — encompassing partial occlusion, font variation, illumination

[7] Plots corresponding to other languages are available in supplementary material

variation, perspective text, multi-oriented text, and text of varying lengths — are meticulously evaluated.

Compared to the current IndicSTR12 dataset, our proposed dataset, is six times larger, resulting in a more extensive collection of unique scene text word images. Additionally, our dataset *IIIT-IndicSTR-word* surpasses existing multi-lingual STR datasets, 1.3 times larger than MLT-19, 2.6 times larger than MLT-17 and 5 times larger than LSVT. It also exceeds the size of existing Latin STR datasets, including IIIT5K-Words, SVT, ICDAR2003, ICDAR2013, ICDAR2015, and SVT Perspective. Our dataset with 250K word images, compares favorably with existing multi-lingual dataset such as MTWI (289K word images). Furthermore, our experimental section explores the potential impact of these differences on the performance and generalization capabilities of models trained on each dataset. This comparative analysis aims to provide a comprehensive understanding of our proposed dataset's distinctive contributions and characteristics within the broader context of existing resources, offering valuable insights for researchers and practitioners alike.

Dataset	Word Images	Language	Features	#Language	#IpL
IIIT5K-Words [16]	5K	English	Reg	1	5K
SVT [29]	725	English	Reg, Blur, LR	1	725
ICDAR2003 [13]	2268	English	Reg	1	2268
ICDAR2013 [8]	5003	English	Reg, SL	1	5003
ICDAR2015 [7]	6545	English	Irr, Blur, Small	1	6545
SVT Perspective [19]	639	English	Irr, PT	1	639
CUTE80 [21]	288	English	Irr, PT, LR	1	288
LSVT [25,26]	50K	Multi-lingual [8]	Irr, MO	2	25K
MLT-19 [17]	191K	Multi-lingual [9]	Irr	10	19.1K
MTWI [5]	289K	Multi-lingual [10]	Irr	2	144.5K
MLT-17 [18]	96K	Multi-lingual [11]	Irr	9	10.6K
Urdu-Text [3]	14K	Urdu	Irr, Noisy	1	14K
IndicSTR12 [14]	27K	Multi-lingual [12]	Irr, LR, Blur, Occ, PT	12	2.25K
IIIT-IndicSTR-Word (Our)	250K	Multi-lingual [13]	Irr, LR, Blur, Occ, PT	10	25K

Table 3. Illustrate comparison of our dataset with various public real scene text recognition datasets. #Language denotes the total number of languages in each dataset, while #IpL represents the average number of word level images per language. Reg, Irr, LR, Occ, Mo and PT, indicate the presence of regular text, irregular text, low resolution images, occlusion, multi oriented text and perspective text, respectively

[8] Chinese and English
[9] Arabic, Bengali, Chinese, Devanagari, English, French, German, Italian, Japanese, and Korean
[10] Chinese and English
[11] Arabic, Bengali, Chinese, English, French, German, Italian, Japanese, and Korean
[12] Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, Telugu, and Urdu
[13] Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu

4 Benchmark Experiments

4.1 Experimental Settings

PARSeq as Baselines: PARSeq, as illustrated in Fig. 7, is a transformer-based model trained utilizing Permutation Language Modeling (PLM) [30,27,20]. Multi-head Attention [28] is a crucial component, denoted as $MHA(q,\ k,\ v,\ m)$, where q, k, v, and m refer to query, key, value, and optional attention mask.

The model follows an encoder-decoder architecture with 12 encoder blocks in the encoder stack and a single block in the decoder. Each of the 12 vision transformer encoder blocks utilizes 1 self-attention MHA module. The image The image $x \in \mathbb{R}^{\mathbb{W}\times\mathbb{H}\times\mathbb{C}}$ is evenly tokenized into $w_p \times w_p$ patches, which are then projected into d - dimensional tokens using an embedding matrix $\boldsymbol{W}^p \in \mathbb{R}^{w_p h_p C \times d}$. Position embeddings are added to the tokens before being sent to the first ViT encoder block. All output tokens $\boldsymbol{z}$ serve as input to the decoder:

$$\boldsymbol{z} = E(x) \in \mathbb{R}^{\frac{WH}{w_p h_p}\times d}$$

The Visio-lingual Decoder is a pre-layerNorm transformer decoder with two MHAs. The first MHA requires position tokens $\boldsymbol{p} \in \mathbb{R}^{(T+1)\times d}$ (with T being the context length), context embeddings $\boldsymbol{c} \in \mathbb{R}^{(T+1)\times d}$, and attention mask $\boldsymbol{m} \in \mathbb{R}^{(T+1)\times(T+1)}$. The position token captures the target position to be predicted and decouples the context from the target position, allowing the model to learn from permutation language modeling. The attention masks vary depending on their use. During training, they are based on permutations, while during inference, a left-to-right look-ahead mask is applied. To enforce the condition that past tokens have no access to future ones, attention masks are used since transformers process all tokens in parallel. In practice, achieving PLM, which theoretically requires the model to train on all $T!$ factorizations, is done by using attention masks to enforce some subset K of $T!$ permutations.

$$\boldsymbol{h}_c = \boldsymbol{p} + MHA(\boldsymbol{p}, \boldsymbol{c}, \boldsymbol{c}, \boldsymbol{m}) \in \mathbb{R}^{(T+1)} \times d$$

The second MHA is employed for image-position attention without any attention mask.

$$\boldsymbol{h}_i = \boldsymbol{h}_c + MHA(\boldsymbol{h}_c, \boldsymbol{z}, \boldsymbol{z} \in \mathbb{R}^{(T+1)} \times d)$$

The last decoder hidden state is used to obtain the output logits $\boldsymbol{y} = Linear(\boldsymbol{h}_{dec} \in \mathbb{R}^{(T+1)\times(S+1)}$, where S is the size of the character set, and an additional 1 is due to the end of sequence token $[\boldsymbol{E}]$. The decoder block can be represented as:

$$\boldsymbol{y} = Dec(\boldsymbol{z}, \boldsymbol{p}, \boldsymbol{c}, \boldsymbol{m}) \in \mathbb{R}^{(T+1)\times(S+1)}$$

Implementation Details: We use synthetic images of IndicSTR12 [14] to train the PARSeq model. To further fine-tune the model for real-world applications, it undergoes additional training, validation, and testing on our dataset, with 70% of the word images allocated for training, 10% for validation, and the remaining 20% for testing in each language.

All PARSeq models are trained on dual-GPU platforms of NVIDIA GeForce GTX 1080 Ti GPU machine using PyTorch DDP, spanning 20-33 epochs with a batch size of 128. We optimize the training process by employing the 1cycle learning rate scheduler [23] in conjunction with the Adam

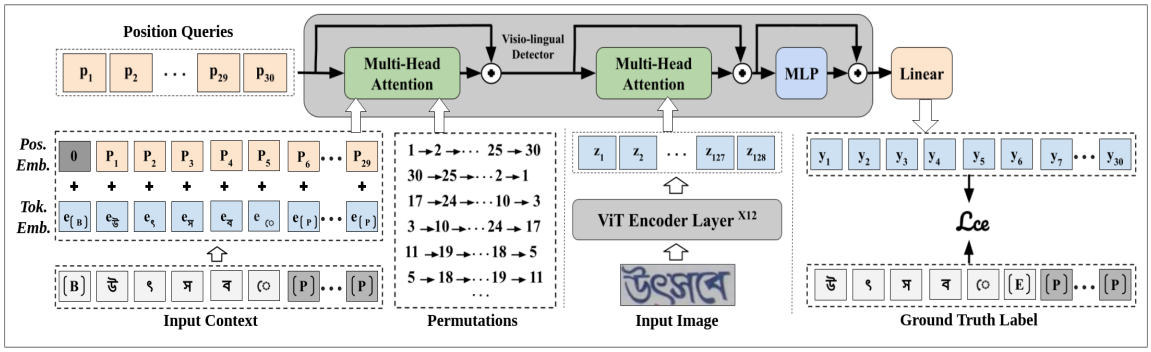


Fig. 7. Presents PARSeq architecture. [B], [E], and [P] stand for beginning-of-sequence (BOS), end-of-sequence (EOS), and padding tokens, respectively. T = 30 or 30 distinct position tokens. L_{ce} corresponds to cross entropy loss.

optimizer [9]. For both PLM and ViTSTR architectures within the PARSeq model, we employ K = 6 permutations with mirroring and an 8 x 4 patch size, respectively. The vocabulary used to create the synthetic dataset determines the maximum label length for the transformer-based PARSeq model. In alignment with community practice [1], we refrain from implementing data augmentation techniques on synthetic datasets.

Training/Testing Details: The PARSeq model is trained on the training set of our dataset. Following this, we evaluate the model's performance on the test sets of the following datasets: *IIIT-IndicSTR-Word* (our), IndicSTR12, MLT-17 (Bengali), MLT-19 (Bengali), and MLT-19 (Hindi).

Evaluation Metrics: We employ two widely acknowledged evaluation metrics: Character Recognition Rate (CRR), also known as Character Error Rate (CER), and Word Recognition Rate (WRR), alternatively referred to as Word Error Rate (WER), to assess the performance of the baseline. The Error Rate (ER) is defined as:

$$ER = (S + D + I)/N. \tag{1}$$

In the context of CER, S represents the number of substitutions, D denotes the number of deletions, I signifies the number of insertions, and N indicates the total number of instances in the reference text. Eq. (1) operates at the character level for CER, while for WER, it operates at the word level. The Recognition Rate (RR) is defined as:

$$RR = 1 - ER. \tag{2}$$

For CRR, Eq. (2) operates at the character level, and for WRR, it functions at the word level.

4.2 Benchmark Results on Word Recognition

Performance on Our Dataset: We utilize synthetic images specific to each language sourced from the IndicSTR12 dataset [14] to pre-train the PARSeq model. Subsequently, we train the PARSeq model using images from the training set of our dataset, followed by an evaluation of the

Fig. 8. Present a few sample visual results on test sets of our dataset. Green, Blue, and Red colored text indicate ground truth, correct prediction, and wrong prediction, respectively.

Language	PARSeq (our)	
	CRR	WRR
Bengali	92.75	85.34
Gujarati	88.12	81.91
Hindi	95.01	87.24
Kannada	87.64	79.27
Malayalam	89.42	80.31
Marathi	94.47	85.50
Oriya	95.13	86.53
Punjabi	91.46	84.27
Tamil	95.63	86.35
Telugu	92.18	84.94

Table 4. Performance of the baseline PARSeq models on test sets of our dataset.

respective test sets for each language. Separate PARSeq models are trained for each language. The results obtained on the test sets of our dataset are presented in Table 4. The table underscores that pre-training the PARSeq model with synthetic images from the IndicSTR12 dataset and then training it with real images from our dataset enhances accuracy in terms of CRR and WRR, owing to the large number of images present in the training set of our dataset. A more extensive training set contributes to improved model performance.

Fig. 8 displays select visual outcomes derived from the test sets of our dataset. The illustrations reveal that PARSeq effectively identifies standard text segments. However, when confronted with multi-oriented or low resolution text, PARSeq struggles to recognize the text accurately.

Dataset	Language	CRNN [4]		STARNet [4]		PARSeq [14]		PARSeq(Our)	
		CRR	WRR	CRR	WRR	CRR	WRR	CRR	WRR
IndicSTR12	Bengali	59.86	48.21	80.26	57.70	83.08	62.04	**89.81**	**82.35**
	Gujarati	52.78	32.81	75.28	46.68	79.75	52.85	**89.81**	**82.35**
	Hindi	78.84	46.56	78.72	46.60	76.01	45.14	**96.11**	**87.93**
	Kannada	78.79	52.43	82.59	59.72	88.64	63.57	**90.46**	**81.02**
	Malayalam	77.94	53.12	84.97	70.09	90.10	68.81	**92.37**	**80.59**
	Marathi	70.79	50.96	83.73	58.65	86.74	63.50	**94.98**	**86.29**
	Oriya	80.39	54.74	86.97	66.30	89.13	71.30	**97.27**	**87.92**
	Punjabi	83.15	68.85	84.93	62.5	92.68	78.70	**94.29**	**86.11**
	Tamil	75.05	59.06	89.69	71.54	87.56	67.35	**94.29**	**86.11**
	Telugu	78.07	58.12	85.52	63.44	92.18	71.94	**94.85**	**85.14**
MLT-17	Bengali	79.98	55.30	85.24	65.73	88.72	71.25	**94.51**	**87.58**
MLT-19	Bengali	82.80	59.51	89.46	71.25	90.10	72.59	**96.28**	**89.32**
	Hindi	86.48	67.90	91.00	75.97	91.80	75.91	**97.12**	**86.45**

Table 5. Performance of the baseline PARSeq models on existing Indic STR datasets.

Performance on Existing Indic STR Datasets We also conducted an additional experiment involving pre-trained PARSeq models, initially trained on synthetic images from the IndicSTR12 dataset [14]. These models were then trained on our dataset and followed by testing the trained models on the existing datasets.

In Experiment-I, the PARSeq model was trained on the training set of our dataset and tested on the validation sets of the IndicSTR12 [14] dataset. Similarly, in Experiment-II, the pre-trained PARSeq model underwent training with the training sets of our dataset and tested on the MLT-17 [18] (Bengali) dataset. In Experiment-III, the pre-trained PARSeq model was trained with the training sets of our dataset and tested on the MLT-19 [17] (Bengali) dataset. Experiment-IV involved the pre-trained PARSeq model being trained with the training sets of our dataset and tested on the MLT-19 [17] (Hindi) dataset. These experiments allowed us to evaluate the adaptability and generalization capabilities of the PARSeq model across different datasets and languages. Table 5 shows the results of these experiments.

Table 5 illustrates the performance enhancements achieved by the PARSeq model through various training strategies. When pre-trained with synthetic images, trained on our dataset, the PARSeq model exhibits improved performance on the corresponding datasets. Specifically, when evaluated on the IndicSTR12 dataset, our PARSeq model demonstrates a notable enhancement, with a 10% increase in CRR and a 15% increase in WRR compared to the PARSeq model presented in [14]. This improvement can be attributed to our dataset's more extensive and diverse nature compared to IndicSTR12, which provides ample data for effectively training the network and thereby boosting its performance. A similar observation is found for the other two datasets, ML-17 and Ml-19.

5 Conclusions

We introduce *IIIT-IndicSTR-Word*, a comprehensive and diverse collection tailored for Indic scene text recognition tasks. Comprising 250K word level images across ten different languages (*Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Oriya, Punjabi, Tamil, and Telugu*), these images

are sourced from roadside scenes captured via a GoPro camera. Notably, the dataset encompasses a broad spectrum of real-world scenarios, including variations in blur, lighting conditions, occlusion, non-standard text orientations, low image resolutions, and perspective distortions. Our study presents benchmark results achieved by applying established architecture for text recognition tasks. Additionally, our experiments demonstrate that training model using our dataset leads to notable improvements in model performance.

Future research avenues could explore end-to-end approaches that integrate text localization and recognition within a unified framework. We eagerly welcome contributions from researchers and developers interested in leveraging this dataset to develop new models and advance the field of Indic scene text recognition.

Acknowledgments

This work is supported by MeitY, Government of India, through the NLTM-Bhashini project.

References

1. Baek, J., Matsui, Y., Aizawa, K.: What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
2. Bušta, M., Patel, Y., Matas, J.: E2E-MLT-an unconstrained end-to-end method for multi-language scene text. In: Asian Conference on Computer Vision (2019)
3. 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 in brief **31**, 105749 (2020)
4. Gunna, S., Saluja, R., Jawahar, C.: Transfer learning for scene text recognition in indian languages. In: International Conference on Document Analysis and Recognition (2021)
5. He, M., Liu, Y., Yang, Z., Zhang, S., Luo, C., Gao, F., Zheng, Q., Wang, Y., Zhang, X., Jin, L.: ICPR 2018 contest on robust reading for multi-type web images. In: International Conference on Pattern Recognition (2018)
6. He, P., Huang, W., Qiao, Y., Loy, C.C., Tang, X.: Reading scene text in deep convolutional sequences. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
7. Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V.R., Lu, S., et al.: ICDAR 2015 competition on robust reading. In: International Conference on Document Analysis and Recognition (2015)
8. Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazan, J.A., De Las Heras, L.P.: ICDAR 2013 robust reading competition. In: International Conference on Document Analysis and Recognition (2013)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2014)
11. Liu, W., Chen, C., Wong, K.Y.K., Su, Z., Han, J.: Star-net: a spatial attention residue network for scene text recognition. In: BMVC (2016)
12. Liu, Z., Li, Y., Ren, F., Goh, W.L., Yu, H.: Squeezedtext: A real-time scene text recognition by binary convolutional encoder-decoder network. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
13. Lucas, S.M., Panaretos, A., Sosa, L., Tang, A., Wong, S., Young, R., Ashida, K., Nagai, H., Okamoto, M., Yamamoto, H., et al.: ICDAR 2003 robust reading competitions: entries, results, and future directions. International Journal of Document Analysis and Recognition **7**, 105–122 (2005)

14. Lunia, H., Mondal, A., Jawahar, C.: IndicSTR12: A dataset for indic scene text recognition. In: International Conference on Document Analysis and Recognition (2023)
15. Mathew, M., Jain, M., Jawahar, C.: Benchmarking scene text recognition in devanagari, telugu and malayalam. In: International Conference on Document Analysis and Recognition (2017)
16. Mishra, A., Alahari, K., Jawahar, C.: Scene text recognition using higher order language priors. In: British Machine Vision Conference (2012)
17. Nayef, N., Patel, Y., Busta, M., Chowdhury, P.N., Karatzas, D., Khlif, W., Matas, J., Pal, U., Burie, J.C., Liu, C.l., et al.: ICDAR 2019 robust reading challenge on multi-lingual scene text detection and recognition—rrc-mlt-2019. In: International conference on document analysis and recognition (2019)
18. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., et al.: ICDAR 2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In: International Conference on Document Analysis and Recognition (2017)
19. Phan, T.Q., Shivakumara, P., Tian, S., Tan, C.L.: Recognizing text with perspective distortion in natural scenes. In: Proceedings of the IEEE International Conference on Computer Vision (2013)
20. Qi, W., Gong, Y., Jiao, J., Yan, Y., Chen, W., Liu, D., Tang, K., Li, H., Chen, J., Zhang, R., et al.: Bang: Bridging autoregressive and non-autoregressive generation with large scale pretraining. In: International Conference on Machine Learning (2021)
21. Risnumawan, A., Shivakumara, P., Chan, C.S., Tan, C.L.: A robust arbitrary text detection system for natural scene images. Expert Systems with Applications **41**(18), 8027–8048 (2014)
22. Saluja, R., Maheshwari, A., Ramakrishnan, G., Chaudhuri, P., Carman, M.: Ocr on-the-go: Robust end-to-end systems for reading license plates and street signs. In: International Conference on Document Analysis and Recognition (2019)
23. Smith, L.N., Topin, N.: Super-convergence: Very fast training of neural networks using large learning rates. In: Artificial intelligence and machine learning for multi-domain operations applications (2019)
24. Su, B., Lu, S.: Accurate scene text recognition based on recurrent neural network. In: Asian Conference on Computer Vision (2015)
25. 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: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
26. Sun, Y., Ni, Z., Chng, C.K., Liu, Y., Luo, C., Ng, C.C., Han, J., Ding, E., Liu, J., Karatzas, D., et al.: ICDAR 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In: International Conference on Document Analysis and Recognition (2019)
27. Tian, C., Wang, Y., Cheng, H., Lian, Y., Zhang, Z.: Train once, and decode as you like. In: Proceedings of International Conference on Computational Linguistics (2020)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)
29. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: International conference on computer vision (2011)
30. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in neural information processing systems (2019)