Unconstrained Camera Captured Indic Offline Handwritten Dataset

Ajoy Mondal[0000−0002−4808−8860] and C V Jawahar[0000−0001−6767−7057]

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

Abstract. This paper presents a diverse compilation of Indic offline handwritten documents. Our dataset comprises 91K handwritten document images captured through unconstrained camera across thirteen Indic languages: *Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Oriya, Punjabi, Tamil, Telugu, and Urdu*, contributed by 1,220 writers. This dataset encompasses 2600K words and includes 566,187 unique words featuring diverse content types, such as alphabetic and numeric. Additionally, we establish a high baseline for the proposed dataset, facilitating evaluation, benchmarking and explicitly focusing on word recognition tasks. Our findings indicate that our dataset is an effective training source for enhancing performance on respective datasets. The code, trained model, dataset, and benchmark results are available at https://cvit.iiit.ac.in/usodi/ucciohd.php.

Keywords: Handwritten text recognition · Indic language · Indic script · camera captured · unconstrained · word recognition · benchmark.

1 Introduction

Advancements in Handwritten Text Recognition (HTR) models underscore the necessity for extensive and meticulously annotated handwritten text recognition datasets. These datasets must exhibit diversity in writing and robust annotation to facilitate reliable performance across real-world applications. While English benefits from established datasets like IAM [16,23] and GNHK [14], which are specifically tailored for offline handwritten text recognition, such comprehensive resources are relatively scarce.

Compared to Latin HTR, the exploration of Indic HTR lags due to the scarcity of annotated resources. India's linguistic diversity presents a unique challenge, with numerous languages and scripts in use [1]. Consequently, amassing substantial handwritten datasets across multiple Indic scripts proves to be arduous and costly. Existing annotated datasets [3,18,7,12] for Indic HTR are limited both in size and breadth. Several initiatives [9,10,11] have endeavored to narrow the disparity between advancements in Latin and Indian languages by introducing a handwritten dataset spanning ten Indian scripts. However, these

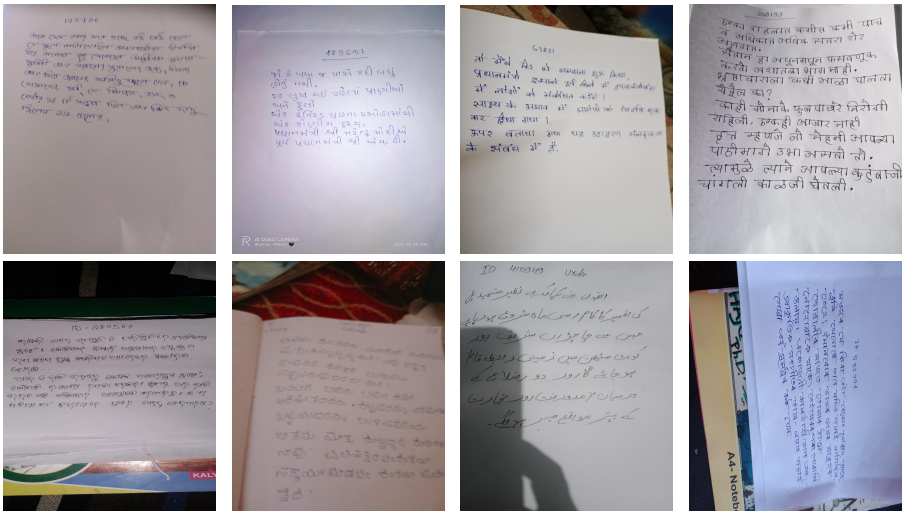


Fig. 1. Shows a few examples of handwritten document images in several Indic languages taken under uncontrolled conditions. This camera- captured images exhibit various characteristics, including blurred text, text with overexposed, perspective text, variation in illumination, unwanted large background, low-resolution text, the text under shadow, oriented text, and others.

datasets typically involve scanned handwritten images captured via flatbed scanners, with writers providing a single word within a box and 20-25 words per page.

With the prevalence of cameras, capturing text in real-world scenarios has become increasingly feasible, allowing us to preserve textual information in pixels. Recently, Zhang *et al.* introduced SCUT-HCCDoc [22], featuring unconstrained camera captured Chinese handwritten documents, while Lee *et al.* [14] presented GNHK, showcasing unconstrained camera captured English handwritten documents. However, such datasets are yet to emerge for offline Indic handwritten documents, highlighting the need for a similar initiative in this domain.

To address this imperative requirement, we present a comprehensive compilation of offline Indic handwritten documents captured in unconstrained settings to facilitate exploration in this domain. Our contributions include the meticulous creation of an innovative dataset explicitly tailored to meet the demands of Indic HTR research. It sets itself apart from existing datasets through a range of key attributes:

– We introduce a dataset, namely *IIIT-Indic-HW-UC*, designed for camera capturing offline Indic handwriting documents in real-world settings (refer Fig.1). It comprises a wide variety of 91K handwritten documents across 13 languages — *Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Oriya, Punjabi, Tamil, Telugu, and Urdu* authored by 1,220 distinct writers all over India and captured by mobile camera. To our

Script	**Language**	**Dataset**	**#W**	**IT**	**#Word**	**#UW**
Bengali	Assamese	IIIT-INDIC-HW-WORDS	-	-	-	-
		IIIT-Indic-HW-UC	80	camera captured	200K	53049
Bengali	Bengali	PBOK [3]	199	flatbed scanner	21K	925
		ROYDB [18]	60	flatbed scanner	17K	525
		CMATERDB2.1 [7]	300	flatbed scanner	18K	120
		IIIT-INDIC-HW-WORDS	24	flatbed scanner	113K	11295
		IIIT-Indic-HW-UC	158	camera captured	200K	34042
Gujarati	Gujarati	IIIT-INDIC-HW-WORDS	17	flatbed scanner	116K	10963
		IIIT-Indic-HW-UC	47	camera captured	200K	45492
Devanagari	Hindi	ROYDB [18]	60	flatbed scanner	16K	1030
		LAW [12]	10	flatbed scanner	27K	220
		IIIT-INDIC-HW-WORDS	12	flatbed scanner	95K	11030
		IIIT-Indic-HW-UC	118	camera captured	200K	31237
Kannada	Kannada	PBOK [3]	57	flatbed scanner	29K	889
		IIIT-INDIC-HW-WORDS	11	flatbed scanner	103K	11766
		IIIT-Indic-HW-UC	71	camera captured	200K	57474
Malayalam	Malayalam	IIIT-INDIC-HW-WORDS	27	flatbed scanner	116K	13401
		IIIT-Indic-HW-UC	137	camera captured	200K	70230
Bengali	Manipuri	IIIT-INDIC-HW-WORDS	-	-	-	-
		IIIT-Indic-HW-UC	101	camera captured	200K	75531
Devanagari	Marathi	IIIT-INDIC-HW-WORDS	-	-	-	-
		IIIT-Indic-HW-UC	95	camera captured	200K	32038
Oriya	Oriya	PBOK [3]	140	flatbed scanner	27K	1040
		IIIT-INDIC-HW-WORDS	10	flatbed scanner	101K	13314
		IIIT-Indic-HW-UC	75	camera captured	200K	34074
Gurmukhi	Punjabi	IIIT-INDIC-HW-WORDS	22	flatbed scanner	112K	11093
		IIIT-Indic-HW-UC	67	camera captured	200K	18264
Tamil	Tamil	TAMIL-DB [20]	50	flatbed scanner	25K	265
		IIIT-INDIC-HW-WORDS	16	flatbed scanner	103K	13292
		IIIT-Indic-HW-UC	78	camera captured	200K	82052
Telugu	Telugu	IIIT-INDIC-HW-WORDS	16	flatbed scanner	120K	12945
		IIIT-Indic-HW-UC	114	camera captured	200K	39514
Nastaliq	Urdu	CENPARMI-U [19]	51	flatbed scanner	19K	57
		IIIT-INDIC-HW-WORDS	8	flatbed scanner	100K	11936
		IIIT-Indic-HW-UC	79	camera captured	200K	27264

Table 1. Illustrates comparison of our proposed dataset with existing Indic handwritten text recognition datasets. #W: indicates the number of writers. #Word: indicates the number of words in the dataset. #UW: indicates the number of unique words. IT: indicates imaging type, either flatbed scanner or camera capture.

knowledge, it is the most extensive and first camera captured dataset for Indic handwritten text recognition (Table 1).

– We offer a baseline model for camera captured Indic handwritten text recognition task (refer Table 3). We employ the cross-dataset analysis method [21] to study generalization aspects comprehensively. It entails training a model on one dataset and assessing its performance on others to gain insights into its adaptability and effectiveness across diverse datasets.

2 Handwritten Datasets in Indic Languages

2.1 Character Level Datasets

Several Indic handwritten character level datasets: DHCD [2], BanglaLekha-Isolated Dataset [8], IITG [5], ISI [6], Kannada-MNIST [17], MNIST-MIX [13] and Urdu [4] are available. The statistics of these datasets are presented in Table 2. These datasets offer a diverse range of handwritten characters from Indic scripts, making them valuable resources for training and evaluating models for character recognition tasks. Researchers and developers can use them to build and test OCR systems, handwriting recognition algorithms, and other applications requiring character level recognition in Indic languages.

Script/Lang.	**Dataset**	**#Image**	**#W**	**#Char.**	**Type**
Devanagari	DHCD	92,000	-	46	character
Bengali	BanglaLekha-Isolated	166,105	-	84	10 numerals 50 basic characters 24 compound characters
Assamese	IITG	12,000	-	52	characters
Devanagari	ISC	22,556	1049	10	numerals
Bengali		12,938	556	10	numerals
Oriya		5970	356	10	numerals
Kannada	Kannada-MNIST	60,000	65	10	numerals
Urdu	MNIST-MIX	45,000	900	50	characters

Table 2. Illustrates statistics of existing character level handwritten datasets in Indic languages. Script/Lang.: indicates script or language, #Image: indicates the number of images, #W: indicates the number of writers, and #Char. indicates the number of characters.

2.2 Word Level Datasets

Several datasets containing word level handwritten samples for various Indic languages, such as PBOK [3], ROYDB [18], CMATERDB2.1 [7], CENPARMI-U [19], LAW [12], TAMIL-DB [20], IIIT-HW-DEV [9], IIIT-HW-TELUGU [10], and IIIIT-INDIC-HW-WORDS [11], are publicly available. Table 1 presents the statistics of these datasets. The table reveals that PBOK, ROYDB, CMATERDB2.1, CENPARMI-U, LAW, and TAMIL-DB datasets contain a larger

number of writers compared to the *IIIT-INDIC-HW-WORDS* dataset. However, the number of words and unique words in the *IIIT-INDIC-HW-WORDS* dataset significantly exceeds those in the PBOK, ROYDB, CMATERDB2.1, CENPARMI-U, LAW, and TAMIL-DB datasets. Table 1 presents the statistics of existing word level handwritten text recognition datasets alongside our newly created dataset. The table reveals that the existing datasets have limitations such as a small number of word level images, lack of writer variations and writing styles, limited linguistic diversity, and samples collected in constrained environments. Additionally, many of these datasets are not publicly available for research. These limitations make it challenging to generalize OCR models for high accuracy on real and diverse handwritten documents. To address these issues, it is necessary to create a dataset with a larger size, diverse writing styles, samples from unconstrained environments, and greater linguistic diversity.

3 IIIT-Indic-HW-UC Dataset

We create a larger, more diverse dataset of offline handwritten documents captured by cameras, known as the *IIIT-Indic-HW-UC* dataset. Compared to the previous version, this dataset features increased language coverage, word count, writer diversity, writing conditions, imaging processes, and ground truth annotation. It includes thirteen major Indic languages: *Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Oriya, Punjabi, Tamil, Telugu, and Urdu*. It consists of 200K word images written by more than 50 writers per language. Writers are required to write corresponding handwritten paragraphs in A4 size white pages, with no constraints on writing style. Handwritten pages are captured using a mobile camera instead of a flatbed scanner. Ground truth annotation is provided at the page level, containing bounding boxes, reading order, textual transcriptions, and the language of all words on the page[1]. We discuss more on it further in the following subsections.

3.1 Data Collection and Annotation

For each language, there are 7K text paragraphs created from the available text corpus[2][3] covering a wide range of topics. Unique id is associated with a paragraph. Each paragraph contains at most 50 words. Users from any geographical area in India have reading and writing capability for any/all 13 Indic languages to be the authentic writers for the data collection. Any authentic writer can write at least 100 and at most 200 paragraphs in a language selected by the writer. The writer must write one paragraph on one A4 sized page. There is no other constraint on the writing. After writing the paragraph(s), the writer takes a picture of the handwritten page with a mobile camera and shares it with

[1] However, in this work, we release only individual word level images and their corresponding ground truth transcriptions.

[2] `https://ufal.mff.cuni.cz/~majlis/w2c/download.html`

[3] `https://corpora.uni-leipzig.de/en?corpusId=ben_wikipedia_2021`

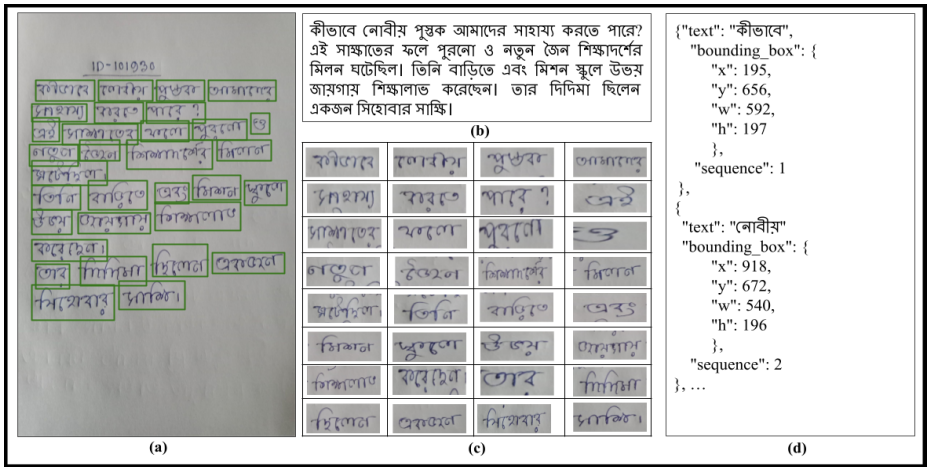


Fig. 2. Illustrates a single Bengali annotated page alongside its standard representation. In (a), a single annotated page from our dataset is depicted. (b) displays the actual text sequence considered as the ground truth. (c) represents all cropped word level images. Lastly, (d) represents textual transcription, bounding box, and sequence of words in a page encapsulated within the JSON file.

us. There are, on average, 50-100 writers for each language to write 7K text paragraphs. Same paragraphs can be written by multiple writers. With the involvement of 1220 writers, we collect a diverse set of 91K handwritten document images corresponding to 91K text paragraphs; each document image is annotated at the page level. The annotation includes complete text paragraphs and bounding boxes, reading order, textual transcriptions, and the language of all words on the page. A sample annotated handwritten document image is depicted in Fig 2. Fig. 2(a) illustrates the ground truth bounding boxes, Fig. 2(b) shows textual transcription for complete document image, Fig. 2(c) depicts cropped word level images, and finally Fig. 2(d) reveals how the ground truth information (textual transcriptions, bounding boxes, and sequences of words) is stored in JSON format.

3.2 Feature and Statistics

Diversity: The handwritten documents contributed by individuals reflecting various age groups, educational backgrounds, and professional experiences across India represent a diverse collection. Capturing these handwritten documents using a mobile camera under unconstrained settings presents numerous challenges, including blurred text, text with overexposed, perspective text, variation in illumination, unwanted large background, low-resolution text, the text under shadow, oriented text, and others. Fig. 1 illustrates a few sample handwritten document images captured under these unconstrained settings. Furthermore, in Fig. 3, we provide several sample word level images of all thirteen languages

Fig. 3. Show several instances of word level images across various languages, written by multiple writers from our dataset.

written by different writers from our dataset, showcasing a wide range of words and highlighting variations in style, image quality, and other aspects.

Since documents are written by various writers all over India, there is enough diversity among documents written by two different writers. Fig. 4 shows a few sample word level images of Hindi written by two different writers: Writer-1 and Writer-21. It highlights that there is still enough variation in writing style and imaging quality between the two writers. Since one writer can write at least 100 and at most 200 pages, for a writer, among document images, there are also enough variations in style and imaging quality because of camera capture. Fig. 5 shows a few sample word level images of Bengali and Hindi languages written by the same writer.

Document Image Resolution Distribution: Writers employ their smartphone cameras to photograph handwritten documents, resulting in variations in the resolution of the captured images. Acknowledging that high-resolution document images offer clear text content, facilitate effective model training, and yield superior performance during testing is crucial. Incorporating document images with diverse resolutions ranging from 1600 $\times$ 720 to 4608 $\times$ 3456 introduces variability in content visibility, thereby enhancing the robustness of the model. Fig. 6(a) highlights the distribution of resolution of handwritten document images by different writers for the Hindi language in our dataset.

Word Level Image Resolution Distribution: Variations in text content and individual writers contribute to differences in the resolution of handwritten word level images. This diversity in word level image resolution improves the model's generalization ability. As depicted in Fig. 6(b), the distribution of resolutions in Hindi word level images highlights the dataset's variability. Most word level

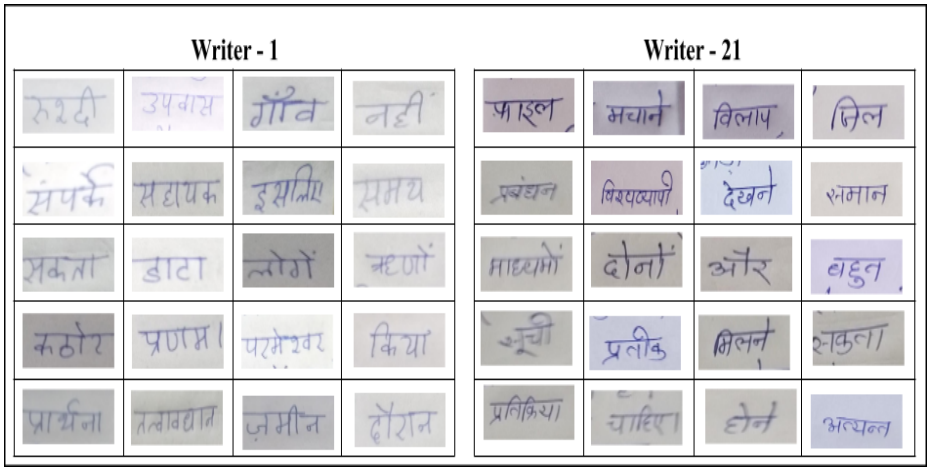


Fig. 4. Presents explicitly sample word level images from just two different writers, namely, writer-1 and writer-21.

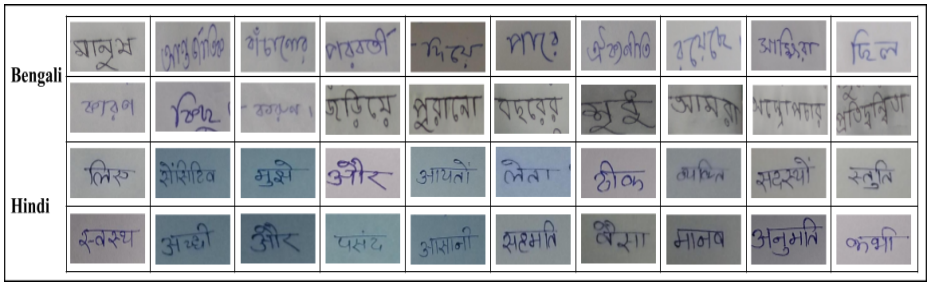


Fig. 5. Show word level images from different document images written by the same writer for Bengali and Hindi languages.

images have a height-to-width ratio between 0.5 and 1.0, so the dataset encompasses word level images of varying resolutions. Including word level images with diverse resolutions enriches the dataset, enabling the model to accommodate a broader range of visual attributes and enhance its performance across various writing styles and conditions.

Contrast of Word Level Images: Word level images are extracted from handwritten document images captured by various mobile cameras, resulting in significant variations in intensities and contrast among the images. To evaluate the recognition ease of each word level image, we adopt the global contrast strategy [15]. Fig. 6(c) illustrates the diverse global contrast levels observed among word level images in Hindi. The figures illustrate contrast levels ranging between 10 and 80. This variability in intensity within word level images adds complexity to the dataset, contributing to the development of robust HTR models. However,

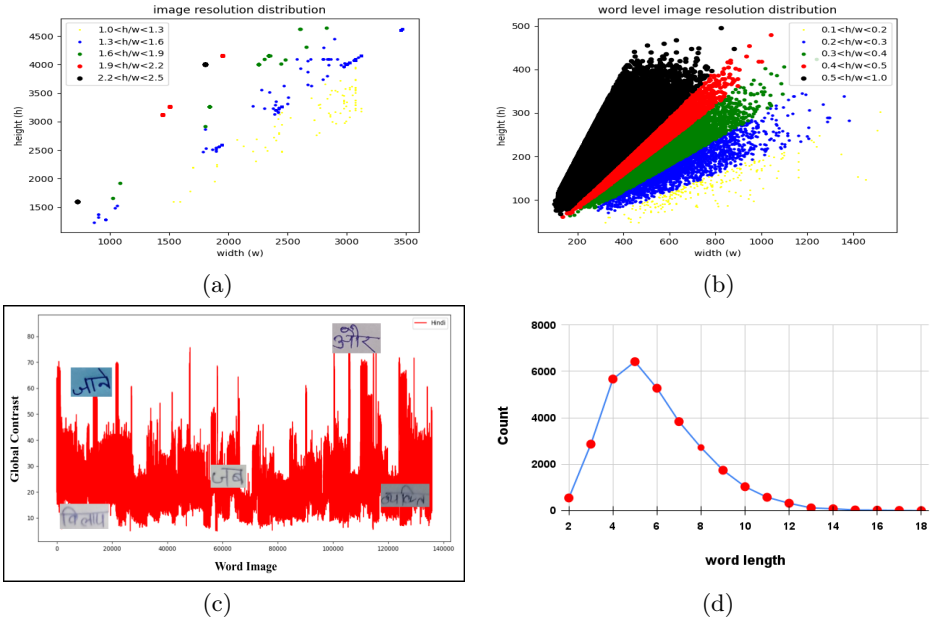


Fig. 6. Shows (a) document image resolution distribution, (b) word level image resolution distribution, (c) varying global contrast among word images, and (d) distribution of word length for the Hindi language in our dataset.

comprehending and utilizing these variations can enhance the adaptability and effectiveness of HTR algorithms across diverse linguistic contexts.

Text Distribution: We compile a dataset of 7K document images for each language, totaling 91K. Additionally, there are 200K word level images per language, resulting in 2600K word level images encompassing unique 566,187 alphabetic and numeric words. Table 1 presents the unique words for each language in our dataset, denoted by the 7^{th} column. For the Hindi language, we include a plot in Fig. 7 that shows the occurrence of unique words in the dataset. The x-axis shows unique words, and the y-axis shows the occurrence of a particular word on a logarithmic scale. This plot demonstrates a long-tail distribution, confirming the diversity of the dataset[4].

Writer Characteristics: Across India, 1,220 individuals have actively contributed to curating handwritten document images, resulting in a diverse dataset encompassing various handwriting styles, camera specifications, scanning methods, and more. Among these contributors, 70% (854 individuals) are female, while the remaining 30% (366 individuals) are male. Within the male cohort, 23

[4] Plots of the occurrence of unique words for other languages in the supplementary material.

individuals are identified as left-handed, with 343 being right-handed. Among the female contributors, 34 individuals are left-handed, while the majority, specifically 820, are right-handed. Notably, a significant portion of the contributors falls within the age range between 20 to 40.

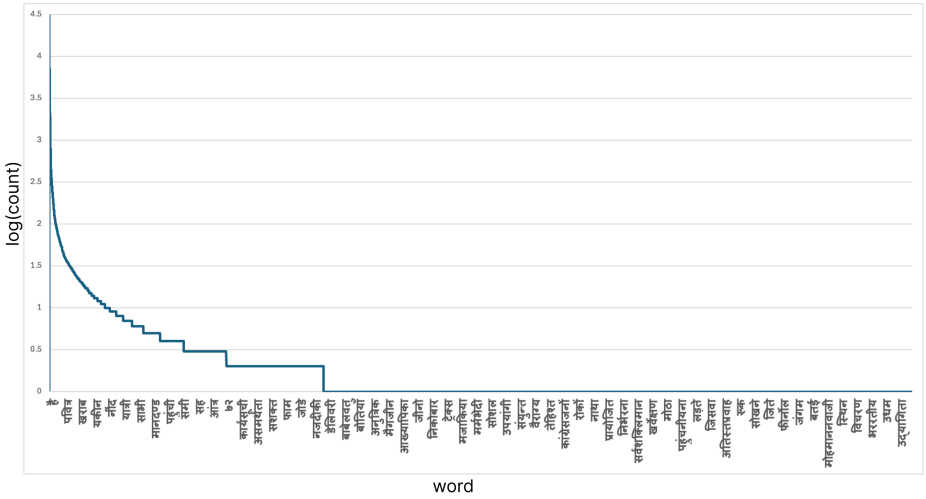


Fig. 7. Shows the distribution of unique Hindi words in the dataset.

Dataset Splits: To furnish an extensive training dataset for deep learning models, our dataset, consisting of 2600K word level images, has been partitioned into 1950K word level images for training, 260K word level images for validation, and 390K word level images for testing. For each language, the dataset includes 200K word level images, among them 75% (i.e., 150K), 10% (i.e., 20K), and 15% (i.e., 30K) word images for training, validation, and test sets, respectively.

Comparison with Existing Datasets: Table 1 comprehensively compares our proposed dataset and existing offline Indic handwritten text recognition datasets, highlighting significant disparities and advantages. Various factors, such as dataset size, diversity in handwriting styles, and the inclusion of diverse texts, are meticulously examined. Our *IIIT-Indic-HW-UC* dataset is twice as large as *IIIT-INDIC-HW-WORDS* regarding the number of word level images, resulting in a more extensive collection of handwritten document images. In contrast to existing datasets where images are scanned using flatbed scanners, our dataset comprises images captured using mobile cameras under unconstrained environments. Camera capture introduces various challenges such as blurred text, overexposed text, perspective distortion, variations in illumination, extensive unwanted backgrounds, low-resolution text, text under shadow, and oriented text, among others, in handwritten document images. Across all

languages in our dataset, the number of writers exceeds that of *IIIT-INDIC-HW-WORDS*, indicating a more diverse range of writing styles. For instance, in the Bengali language, the PBOK and CMATERDB2.1 datasets boast more writers (199 and 300, respectively) compared to our dataset, with 158 writers. However, our dataset contains a more significant number of unique words (34042) than the PBOK and CMATERDB2.1 datasets, which have 925 and 25 unique words, respectively. Similarly, in the Oriya language, while the PBOK dataset has more writers (140) than our dataset (75), our dataset surpasses PBOK in terms of the number of unique words (34074 versus 1040). These distinguishing characteristics render our dataset larger, covering major Indic languages and offering diverse word level images compared to existing datasets.

4 Benchmark Experiments

4.1 Experimental Settings

Baseline: We utilize the network architecture proposed by Gongidi *et al.* [11], depicted in Fig. 8, as the baseline for our experiment. This network consists of four main modules: the Transformation Network (TN), Feature Extractor (FE), Sequence Modeling (SM), and Predictive Modeling (PM). The Transformation Network comprises six plain convolutional layers with 16, 32, 64, 128, 128, and 128 channels, each followed by a max-pooling layer of size 2×2 and a stride of 2. The Feature Extractor module adopts the ResNet architecture, while the Sequence Modeling module employs a 2-layer Bidirectional LSTM (BLSTM) with 256 hidden neurons in each layer. Finally, the Predictive Modeling module utilizes Connectionist Temporal Classification (CTC) for character decoding and recognition by aligning the feature sequence with the target character sequence. Further details can be found in [11].

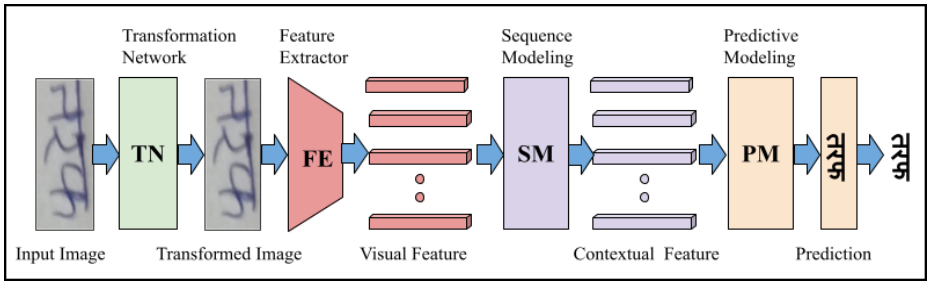


Fig. 8. Shows text recognition through the baseline pipeline.

Implementation Details: The baseline model is trained using a single NVIDIA GeForce GTX 1080 Ti GPU. Word level images are resized to dimensions of 96

$\times$ 256 during pre-processing. Stochastic Gradient Descent (SGD) is utilized with the Adadelta optimizer, employing a learning rate of 0.001, a batch size of 64, and a fixed momentum of 0.09. Upon acceptance of the paper, both the trained model and dataset will be released to the public.

Training/Testing Details: The baseline model is trained on the training sets of our *IIIT-Indic-HW-UC*, *IIIT-INDIC-HW-WORDS*, and a combination of these two datasets. Following training, we evaluate the performance of the baseline model on the respective test sets of *IIIT-Indic-HW-UC* and *IIIT-INDIC-HW-WORDS* datasets.

Evaluation Metrics: We utilize two widely recognized evaluation metrics, namely, Character Recognition Rate (CRR) (alternatively Character Error Rate, CER) and Word Recognition Rate (WRR) (alternatively Word Error Rate, WER), to evaluate the performance of the baseline. Error Rate (ER) is defined as:

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

where 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 reference text. In the context of CER, Eq. (1) operates at the character level, and while of WER, Eq. (1) 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.

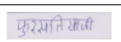
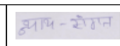
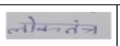
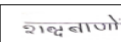
Language	Trained on	Tested on			
		IIIT-INDIC-HW-WORDS		IIIT-Indic-HW-UC	
					
Hindi	IIIT-INDIC-HW-WORDS	फुरसतियाजी फुरसतियाजी	न्याय-संगत न्याय-संगन	लोकतंत्र लोकतंत्र	शब्दबाणों शन्दबाणों
	IIIT-Indic-HW-UC	फुरसतियाजी फुरसतियाजी	न्याय-संगत आययसंगत	लोकतंत्र लोकतंत्र	शब्दबाणों शब्दबाणों
	IIIT-INDIC-HW-WORDS and IIIT-Indic-HW-UC	फुरसतियाजी फुरसतियाजी	न्याय-संगत न्याय-संगत	लोकतंत्र लोकतंत्र	शब्दबाणों शब्दबाणों

Fig. 9. Visual results obtained by baseline for Hindi language. Ground truth text is highlighted in Blue color. Correctly recognized text is highlighted in Black color. Wrongly recognized text is highlighted in Red color.

Language	**Training Set**	**Test Dataset**			
		IIIT-INDIC-HW-WORDS		*IIIT-Indic-HW-UC*	
		CRR	**WRR**	**CRR**	**WRR**
Assamese	*IIIT-Indic-HW-UC*	-	-	90.98	85.15
Bengali	*IIIT-INDIC-HW-WORDS*	95.72	84.29	79.66	51.51
	IIIT-Indic-HW-UC	85.99	53.82	96.60	89.10
	Both	**96.28**	**84.58**	**96.69**	**89.27**
Gujarati	*IIIT-INDIC-HW-WORDS*	96.16	81.41	81.64	54.14
	IIIT-Indic-HW-UC	85.41	56.96	97.64	88.24
	Both	**97.39**	**87.53**	**98.01**	**88.74**
Hindi	*IIIT-INDIC-HW-WORDS*	96.08	83.09	72.27	40.81
	IIIT-Indic-HW-UC	89.66	63.65	93.88	84.69
	Both	**96.64**	**85.20**	**93.96**	**85.06**
Kannada	*IIIT-INDIC-HW-WORDS*	98.69	92.36	84.52	55.36
	IIIT-Indic-HW-UC	87.96	61.50	97.90	91.49
	Both	**98.83**	**92.41**	**98.15**	**91.59**
Malayalam	*IIIT-INDIC-HW-WORDS*	98.01	89.78	74.97	32.07
	IIIT-Indic-HW-UC	86.77	55.37	96.13	86.74
	Both	**98.16**	**89.81**	**97.05**	**87.07**
Manipuri	*IIIT-Indic-HW-UC*	-	-	90.99	83.38
Marathi	*IIIT-Indic-HW-UC*	-	-	96.32	86.67
Oriya	*IIIT-INDIC-HW-WORDS*	96.02	80.82	84.72	52.35
	IIIT-Indic-HW-UC	85.64	53.13	94.09	80.43
	Both	**96.47**	**82.91**	**94.89**	**81.34**
Punjabi	*IIIT-INDIC-HW-WORDS*	94.87	81.63	69.66	46.60
	IIIT-Indic-HW-UC	83.32	50.11	94.82	87.42
	Both	**96.59**	**86.63**	**94.95**	**87.54**
Tamil	*IIIT-INDIC-HW-WORDS*	98.42	92.19	78.27	41.24
	IIIT-Indic-HW-UC	94.65	73.55	95.31	83.93
	Both	**98.71**	**92.36**	**96.29**	**84.06**
Telugu	*IIIT-INDIC-HW-WORDS*	95.42	76.02	76.73	35.05
	IIIT-Indic-HW-UC	94.50	71.46	93.43	74.34
	Both	**97.50**	**83.60**	**93.68**	**74.84**
Urdu	*IIIT-INDIC-HW-WORDS*	93.50	75.89	71.25	43.94
	IIIT-Indic-HW-UC	75.47	47.07	93.21	82.18
	Both	**96.11**	**85.01**	**94.59**	**82.93**

Table 3. Quantitative results on different Indic handwritten datasets. Bold value indicates the best results.

4.2 Benchmark Results on Word Level Text Recognition

The performance evaluation results of our baseline model on offline Indic handwritten datasets are presented in Table 3. The table illustrates that when the model is trained on *IIIT-INDIC-HW-WORDS* and tested on both *IIIT-INDIC-*

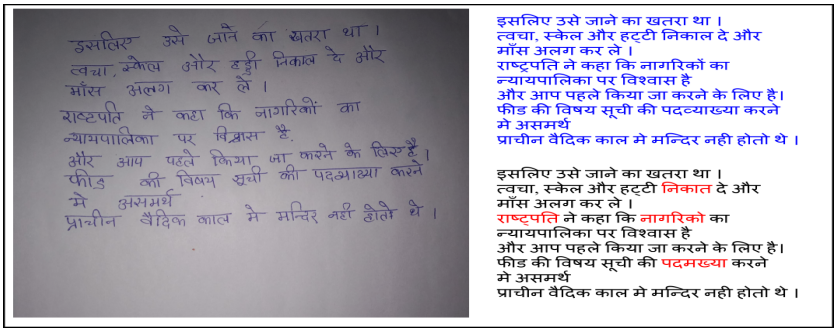


Fig. 10. Page level visual results obtained by baseline for Hindi language. Ground truth text is highlighted in Blue color. Correctly recognized text is highlighted in Black color. Wrongly recognized text is highlighted in Red color.

HW-WORDS and *IIIT-Indic-HW-UC*, it achieves notably higher accuracy on *IIIT-INDIC-HW-WORDS* compared to *IIIT-Indic-HW-UC* due to the domain gap (flatbed vs. camera captured and constrained vs. unconstrained) between these datasets. Similarly, when the model is trained on *IIIT-Indic-HW-UC* and tested on both *IIIT-INDIC-HW-WORDS* and *IIIT-Indic-HW-UC*, it achieves better results on *IIIT-Indic-HW-UC* than *IIIT-INDIC-HW-WORDS*, again due to the domain gap between the datasets. We also noticed that when the model is trained on both *IIIT-INDIC-HW-WORDS* and *IIIT-Indic-HW-UC* and then tested on these two datasets, it achieved the highest performance (indicated by bold values in Table 3) for all languages across both datasets. We noticed that when the model was trained with *IIIT-Indic-HW-UC* and tested on *IIIT-INDIC-HW-WORDS*, it achieved better WRR and CRR compared to when the model was trained with *IIIT-INDIC-HW-WORDS* and tested on *IIIT-Indic-HW-UC*. The unique properties of the *IIIT-Indic-HW-UC* dataset, such as unconstrained and camera captured images, contribute to the model's generality, enabling better performance even on the *IIIT-INDIC-HW-WORDS* dataset, which comprises images captured by a flatbed scanner in a constrained environment. This suggests that the *IIIT-Indic-HW-UC* dataset is more diverse and makes the model more generic than the *IIIT-INDIC-HW-WORDS* dataset. Figs. 9 and 10 display visual results for the Hindi language at both word and page levels, respectively. In these figures, ground truth text is highlighted in blue, correctly recognized text in black, and wrongly recognized text in red.

4.3 APIs and Web-based Applications

We develop APIs for handwritten page recognition models across 13 languages and create a web-based application that integrates these APIs to digitize handwritten documents in Indic languages. Fig. 11 illustrates the steps for using our web-based APIs to digitize Indic handwritten documents. Users can upload a

handwritten document image, select the language, choose the OCR model version and layout version, and then execute the process to obtain the OCR output.

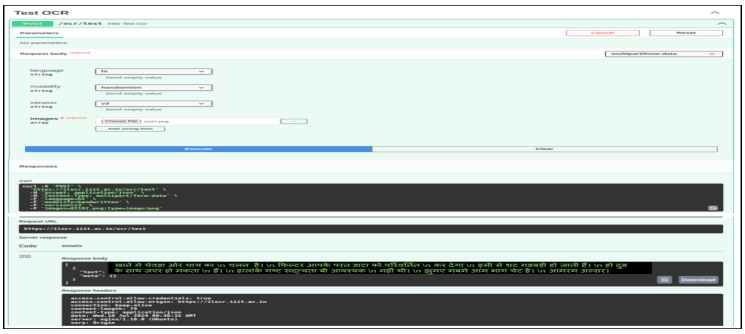


Fig. 11. Shows screen shot of our web-based APIs to digitize Indic handwritten documents.

5 Conclusion

We have introduced a large and diverse dataset called *IIIT-Indic-HW-UC* for Indic offline handwritten text recognition. This dataset contains camera captured images of handwritten documents in thirteen Indic languages: *Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Manipuri, Marathi, Oriya, Punjabi, Tamil, Telugu, and Urdu*. We gathered these samples from various regions in India. The dataset includes 91K text paragraphs written by 1,220 writers, offering a wide range of content. We identified 26K instances of words within these images, including alphabetic and numeric text. Among these, 566,187 instances are unique. Our paper presents benchmark results for text recognition using established architectures, showing that training models with our dataset improves performance on existing offline handwritten datasets. We suggest that future research could explore end-to-end approaches integrating localization and recognition. We invite contributions from researchers and developers to explore new models using our dataset.

Acknowledgment

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

References

1. Census 2011. `https://censusindia.gov.in/2011-Common/CensusData2011.Html`

2. Acharya, S., Pant, A.K., Gyawali, P.K.: Deep learning based large scale handwritten devanagari character recognition. In: SKIMA (2015)
3. Alaei, A., Pal, U., Nagabhushan, P.: Dataset and ground truth for handwritten text in four different scripts. IJPRAI **26**(04), 1253001 (2012)
4. Ali, H., Ullah, A., Iqbal, T., Khattak, S.: Pioneer dataset and automatic recognition of urdu handwritten characters using a deep autoencoder and convolutional neural network. SN Applied Sciences **2**, 1–12 (2020)
5. Baruah, U., Hazarika, S.M.: A dataset of online handwritten assamese characters. J. Inf. Process. Syst. **11**(3), 325–341 (2015)
6. Bhattacharya, U., Chaudhuri, B.: Databases for research on recognition of handwritten characters of indian scripts. In: ICDAR (2005)
7. Bhowmik, S., Malakar, S., Sarkar, R., Basu, S., Kundu, M., Nasipuri, M.: Off-line bangla handwritten word recognition: a holistic approach. Neural Computing and Applications **31**, 5783–5798 (2019)
8. Biswas, M., Islam, R., Shom, G.K., Shopon, M., Mohammed, N., Momen, S., Abedin, M.A.: Banglalekha-isolated: A comprehensive bangla handwritten character dataset. arXiv preprint arXiv:1703.10661 (2017)
9. Dutta, K., Krishnan, P., Mathew, M., Jawahar, C.: Offline handwriting recognition on devanagari using a new benchmark dataset. In: DAS (2018)
10. Dutta, K., Krishnan, P., Mathew, M., Jawahar, C.: Towards spotting and recognition of handwritten words in indic scripts. In: ICFHR (2018)
11. Gongidi, S., Jawahar, C.: IIIT-INDIC-HW-Words: A dataset for indic handwritten text recognition. In: ICDAR (2021)
12. Jayadevan, R., Kolhe, S.R., Patil, P.M., Pal, U.: Database development and recognition of handwritten devanagari legal amount words. In: ICDAR (2011)
13. Jiang, W.: Mnist-mix: a multi-language handwritten digit recognition dataset. IOP SciNotes **1**(2), 025002–025010 (2020)
14. Lee, A.W., Chung, J., Lee, M.: Gnhk: A dataset for english handwriting in the wild. In: ICDAR (2021)
15. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR (2014)
16. Marti, U.V., Bunke, H.: The IAM-database: an english sentence database for offline handwriting recognition. IJDAR **5**, 39–46 (2002)
17. Prabhu, V.U.: Kannada-MNIST: A new handwritten digits dataset for the kannada language. arXiv preprint arXiv:1908.01242 (2019)
18. Roy, P.P., Bhunia, A.K., Das, A., Dey, P., Pal, U.: Hmm-based indic handwritten word recognition using zone segmentation. Pattern recognition **60**, 1057–1075 (2016)
19. Sagheer, M.W., He, C.L., Nobile, N., Suen, C.Y.: A new large urdu database for off-line handwriting recognition. In: ICIP (2009)
20. Thadchanamoorthy, S., Kodikara, N., Premaretne, H., Pal, U., Kimura, F.: Tamil handwritten city name database development and recognition for postal automation. In: ICDAR (2013)
21. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR (2011)
22. Zhang, H., Liang, L., Jin, L.: Scut-hccdoc: A new benchmark dataset of handwritten chinese text in unconstrained camera-captured documents. Pattern Recognition **108**, 107559 (2020)
23. Zimmermann, M., Bunke, H.: Automatic segmentation of the IAM off-line database for handwritten english text. In: ICPR. vol. 4, pp. 35–39 (2002)