Bridging the Gap in Resource for Offline English Handwritten Text Recognition

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

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

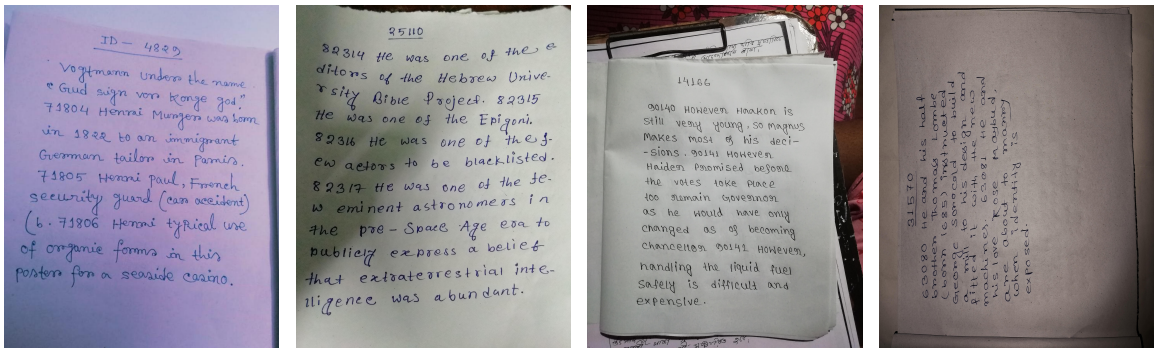


Fig. 1. Examples from our *IIIT-HW-English-Word*, an offline camera-captured English handwritten dataset. Best viewed in color and zoomed-in.

Abstract. The practical applications of Handwritten Text Recognition (HTR) have flourished with many successful commercial APIs, solutions, and diverse use cases. Despite the availability of numerous industrial solutions, academic research in HTR, particularly for English, has been hindered by the scarcity of publicly accessible data. To bridge this gap, this paper introduces *IIIT-HW-English-Word*, a large and diverse collection of offline handwritten English documents. This dataset comprises unconstrained camera-captured images featuring 20,800 handwritten documents crafted by 1,215 writers. Within this dataset, covering 757,830 words, we identify 174,701 unique words encompassing a variety of content types, such as alphabetic, numeric, and stop-words. We also establish a baseline for the proposed dataset, facilitating evaluation and benchmarking, explicitly focusing on word recognition tasks. Our findings suggest that our dataset can effectively serve as a training source to enhance performance on respective datasets. The code, dataset, and benchmark results are available at `https://cvit.iiit.ac.in/usodi/bgroehtr.php`.

Keywords: Handwritten text recognition, English, offline, unconstrained, camera-captured, word recognition, and benchmark.

Dataset	#Pages	#Writers	Imaging Type	#Words	#Unique Words
IAM [17,8]	1,539	657	Flatbed-scanned	115,320	10,480
GNHK [7]	687	-	Camera-captured	39,026	12,341
IIIT-HW-English-Word (Our)	20,800	1,215	Camera-captured	757,830	174,701

Table 1. Illustrates comparison of our dataset with existing offline English handwritten text recognition datasets.

Dataset	Unique Word				
	#Alphabetic	#Numeric	#Stop-word	#Other	#Total
IAM	9,103	116	140	1,121	10,480
GNHK	6,649	250	141	4,194	12,341
IIIT-HW-English-Word (Our)	66,324	97,916	137	10,324	174,701

Table 2. Demonstrates a comparative analysis between our dataset and existing offline English handwritten text recognition datasets, focusing on the diversity and uniqueness of words across different categories.

1 Introduction

The progress of Handwritten Text Recognition (HTR) models demands the availability of a large and high-quality English handwritten text recognition dataset that is diverse and well-annotated. This dataset should be representative enough to ensure effective generalization in real-world scenarios. Presently, there are only a handful of widely used datasets, such as IAM [17,8], and GNHK [7], dedicated to offline handwritten text recognition. Although these datasets provide crucial data for handwritten text, they present a constraint regarding training recognition models, particularly deep architectures like transformer-based models, which may find the data insufficient for effective training.

Hence, there is an escalating need for a more expansive and varied dataset designed explicitly for offline English handwritten images captured in natural settings. This demand has arisen to address the evolving research requirements in HTR. Establishing such a dataset is paramount for pushing the boundaries of research in this domain fostering advancements in HTR technologies. We introduce *IIIT-HW-English-Word*, a comprehensive and diverse compilation of offline handwritten English documents captured in unconstrained settings to fulfill this imperative need. We present noteworthy contributions in our pursuit of facilitating an exploration of this domain. Foremost among these is the meticulous curation of the innovative dataset, deliberately tailored to meet the demands of HTR research. It sets itself apart from existing datasets through a range of key attributes:

– Introducing a dataset, designed for capturing offline English handwriting in authentic, real-world settings. This dataset comprises 20,800 document images, encapsulating a wide variety of 757,830 words authored by 1,215 distinct writers, refer to Fig. 1. For an overview and a comparison with existing datasets such as IAM and GNHK, refer to Table 1. The presented statistics demonstrate that compared to current datasets, ours is notably 13 times and 30 times larger than IAM and GNHK concerning page images, leading to a more extensive collection of unique words in the dataset.

- Within the dataset encompassing a total of 757,830 words, there exist 174,701 unique words, including alphabetic, numeric, stop-words, and other categories. Among these 174,701 unique words, a breakdown reveals that 66,324 are uniquely alphabetic, 97,916 are uniquely numeric, 137 are uniquely stop-words, and the remaining 10,324 falls into the "other" category, encompassing special symbols, characters, and combinations of alphabets and numerals. Table 2 highlights the diversity and uniqueness of words across various categories in our dataset, surpassing the existing two datasets.
- Furthermore, we establish a baseline for evaluating the proposed dataset in the context of word recognition tasks. The significance of datasets in shaping the generalization and assessing the difficulty of various algorithms cannot be overstated, as highlighted by previous research [16]. We employ the cross-dataset analysis method [14] to study these aspects comprehensively. It involves training a model on one dataset and testing its performance on others, providing insights into its adaptability and performance across diverse datasets (see Table 3).

2 Related Work

Offline handwriting recognition is challenging due to the inherent diversity and variability in handwriting styles, making it difficult to build accurate and robust recognition systems. Researchers and developers in document analysis and handwriting recognition have made significant efforts to overcome these challenges. One crucial aspect of advancing the area is the creation of benchmark datasets that serve as a standardized evaluation platform for various handwriting recognition methods. These datasets usually contain many handwritten samples covering multiple writing styles and content types.

Latin Handwritten Datasets: Several widely recognized datasets are crucial for advancing offline handwriting recognition tasks for Latin scripts such as English and French. These datasets, each with unique characteristics, contribute significantly to evaluating and benchmarking recognition systems. Here are some notable datasets: IAM (English) [17,8], IUPR (English) [2], IRONOFF (English) [15], GHNK (English) [7], Deepwriting (English) [1], Belfort (English) [13], and RIMES (French) [6].

The IAM dataset encompasses English text authored by 657 writers, comprising 1,539 paragraphs, 5,685 sentences, 13,353 lines, and 115,320 words. The IUPR dataset is widely recognized and features an extensive collection of handwritten text samples. On the other hand, the IRONOFF dataset focuses on isolated characters, digits, and cursive words written by French writers. The GNHK dataset comprises camera-captured images of English handwritten text from diverse global regions, totaling 687 document images. It encompasses 9,363 lines, 39,026 words, and 172,936 characters. On the other hand, the Deepwriting dataset combines IAM-OnDB with fresh samples, featuring 85,181 words and 406,956 characters handwritten by 294 distinct authors. The Belfort dataset comprises minutes from the Belfort municipal council spanning 1790 to 1946, featuring 24,105 text-line images. Each line image is associated with up to four transcriptions. On the other hand, the RIMES dataset, a widely used resource for offline handwritten text recognition in French, encompasses 5,605 real mail. This dataset encompasses 250,000 words, showcasing diverse writing styles and content types.

Latin Historical Handwritten Datasets: Several historical handwritten datasets in Latin script, including Bentham [10], George Washington [4], Saint Gall [3], Digital Peter [9], READ [11], and SIMARA [12], are available for research purposes.

The Bentham dataset contains over 6000 documents written by the famous English philosopher and reformer Jeremy Bentham. It is a valuable source of information containing over 100K running words with extensive lexicons. On the other hand, the George Washington database was created at the Library of Congress using George Washington Papers dating back to the eighteenth century. It comprises 20 pages, 656 text lines, 4894 words, 1471 classes of words, and 82 letters, providing a glimpse into the past. The Saint Gall dataset is another intriguing historical document, consisting of 60 pages with 1410 text lines and 11597 words of manuscripts written in Latin script in the 9th century. Lastly, the Digital Peter dataset comprises Peter the Great's manuscripts annotated for segmentation and text recognition, containing a staggering 9694 images and text files corresponding to lines with 265788 symbols and approximately 50998 words in historical documents. The READ dataset comprises about 30000 pages written in Early Modern German, and the ground truth in this set is provided in PAGE format with line-level annotations in the PAGE files. On the other hand, the SIMARA corpus consists of 5393 finding aids from six series spanning the 18th-20th centuries, which can be used to extract information from historical handwritten documents.

We could highlight the significant focus of HTR research on historical documents, yet there exists a vast array of practical and immediate applications for HTR in modern handwritten documents. Our emphasis lies within this domain.

3 IIIT-HW-English-Word Dataset

From an extensive English text corpus[1], a curated compilation of 20,800 text paragraphs has been meticulously crafted. Each paragraph receives a unique ID and is constrained to a maximum of 50 words. Users worldwide who are proficient in English reading and writing are assigned up to 30 paragraphs to transcribe onto standard A4-sized pages, allowing unrestrained and creatively expressive writing. Following this, writers are invited to scan the written pages using any scanning app or capture images using a mobile camera. With the involvement of over 1,215 writers, we collect a diverse set of 20,800 handwritten document pages, each page being annotated in a bounding box and text transcript at word level on a page. In Fig.2, a sample annotation is showcased. Fig.2(a) illustrates the ground truth bounding boxes along with their corresponding text transcriptions, (b) displays the actual text sequences, and (c) reveals how the ground truth information is stored in JSON format.

After creating a corpus, we need effective and efficient methods to access the dataset. The recognizers may require the data in a specific format. Thus, a large dataset calls for a standard representation independent of the script. With the generation of extensive data, there is a need for an efficient access mechanism. The data should be organized and stored in a structured way for efficient access. A standard is required to ensure the ease of access by a spectrum of communities that may need the data. JSON provides an efficient method of data storage. It is easy to write applications on a standard dataset. Even updation and changes to the data can be done quickly in a JSON storage standard. All communities of the world accept JSON for data representation. For a handwritten page or document, a JSON contains word bounding boxes and their corresponding textual transcriptions. A set of standard application program interfaces (APIs) are required for adequate access to the datasets. The annotation data (the text and word image) is required for the development and performance evaluation of recognizers.

[1] `https://wortschatz.uni-leipzig.de/en/download/English`

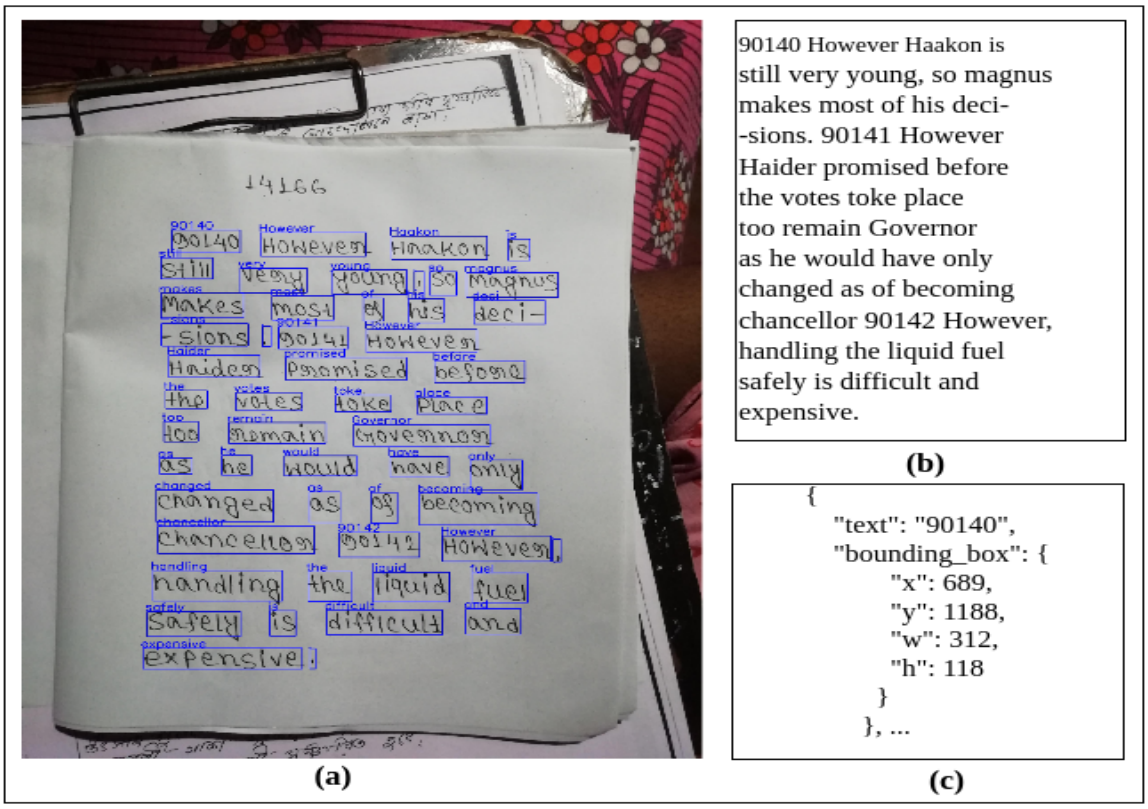


Fig. 2. Showcases a single annotated handwritten document page alongside its standard representation. In (a), a single annotated page from our dataset is depicted. (b) Displays the actual text sequence considered as the ground truth. Finally, (c) represents the content encapsulated within the JSON file.

3.1 Dataset Feature and Statistics

Diversity: As the handwritten document pages are contributed by individuals spanning various age groups, educational backgrounds, and professional experiences under unconstrained settings across India, the resulting collection is characterized by its diversity. The process of capturing these handwritten document pages using a mobile camera under unconstrained settings introduces several challenges, encompassing (i) varying illumination, (ii) shadows, (iii) extensive unwanted background, (iv) presence of irrelevant background text, (v) fluctuations in orientation, vi) low resolution, and (vii) skewed page alignment. A few sample handwritten images captured under unconstrained settings are depicted in Fig. 3. Additionally, we provide several sample word level images from our dataset in Fig.4; a diverse range of words is depicted, showcasing variations in style, imaging, quality, and other aspects.

Page Image Resolution Distribution: Writers utilize their smartphone cameras to capture images of handwritten document pages, leading to variations in the resolution of the captured page images. Acknowledging that high-resolution document images offer clear text content, facilitate

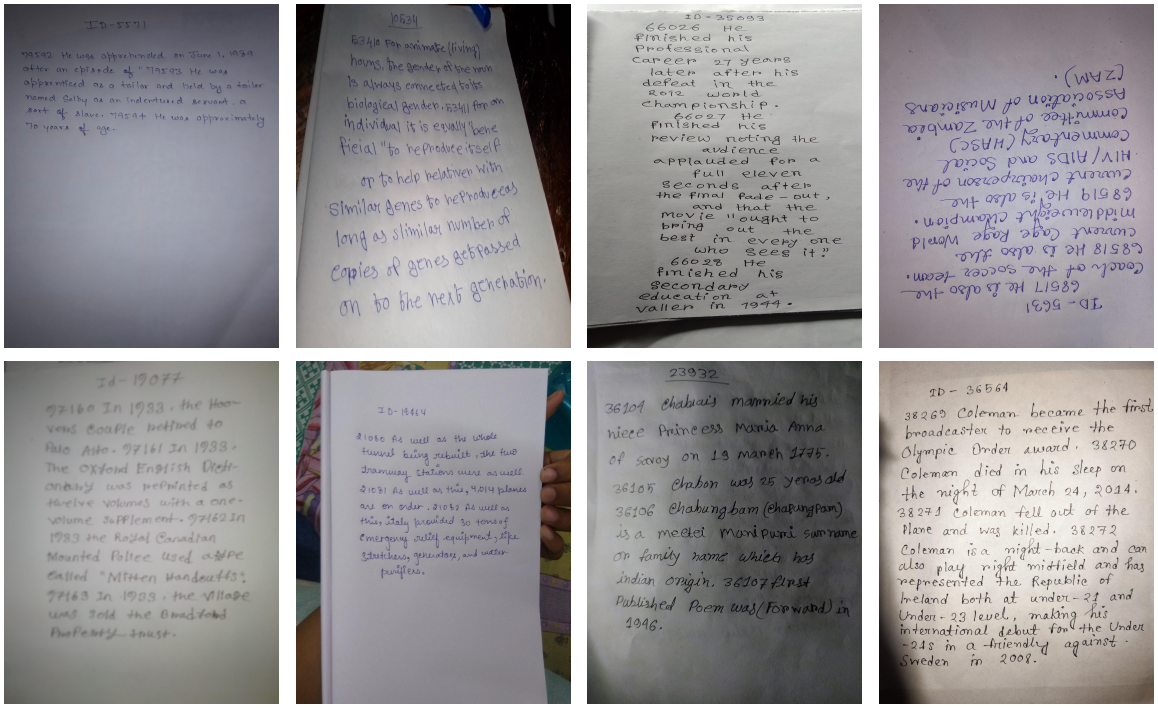
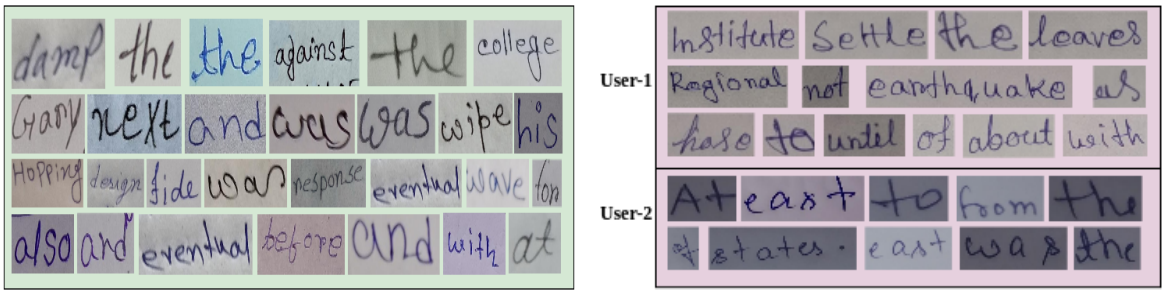


Fig. 3. Shows examples of handwritten document page images taken under uncontrolled conditions. These camera-captured images exhibit various characteristics, including blurred text, text with unwanted large backgrounds, oriented text, variations in illumination, and text with overexposure, among others.

effective model training, and yield superior performance during testing is crucial. Including page images with diverse resolutions introduces variability in content visibility, contributing to the robustness of the model. Fig. 5(a) displays the distribution of resolutions in page images, showcasing a substantial number of images ranging from 2500×1500 to 4000×3000. This visualization offers insights into the dataset's inherent variability. This diversity in resolution further enhances the adaptability and generalization capabilities of the model.

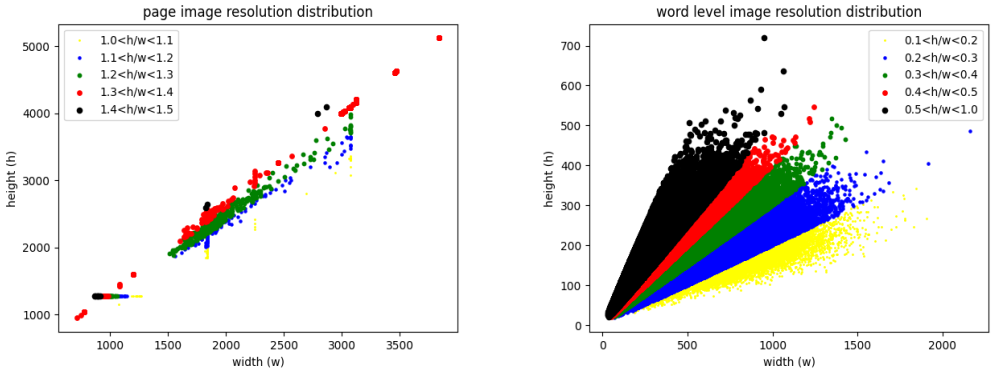
Word Level Image Resolution Distribution: The diversity in both the text content and the individual writers contributes to variations in the resolution of handwritten words. This diversity in word image resolution plays a crucial role in enhancing the model's generalization capabilities. Fig. 5(b) provides a visual representation of the distribution of resolutions in word level images, showcasing the range of variability present in the dataset. Most word images have a height-to-width ratio between 0.1 and 0.25, resulting in word images having various resolution. Including words with varying resolutions enriches the dataset, enabling the model to adapt to a broader spectrum of visual characteristics and improving its performance across different writing styles and conditions.

Text Distribution: Within the compilation of 20,800 document page images, 757,830 instances of words have been identified, covering a spectrum of alphabetic, numeric, stop-words, and text featuring combinations of special symbols, alphabets, and numeric characters. Among these occurrences,



(a) (b)

Fig. 4. A few examples of word level images from our dataset: (a) showcases sample word level images from all users, while (b) presents explicitly sample word level images from just two users, namely, user-1 and user-2.



(a) (b)

Fig. 5. Left Image: page image resolution distribution and Right Image: word level image resolution distribution. For page images, the majority exhibit a height-to-width ratio ranging from 1.29 to 1.33. Conversely, most word images have a height-to-width ratio between 0.1 and 0.25.

174,701 are deemed unique, encompassing alphabetic, numeric, stop-word, and other categories. A more detailed breakdown of these 174,701 unique words reveals that 66,324 are uniquely alphabetic, 97,916 are uniquely numeric, 137 are uniquely stop-words, and the remaining 10,324 fall into other categories. On average, each page image contains approximately 37 instances of words. Fig. 6 visually presents the word cloud distribution of unique alphabetic words in the dataset. The word 'also' is most occurring. While 'one', 'first', 'people', 'however' are frequently occurring words. Visualize the distribution of the dataset's top 70 most frequently occurring alphabetic words using Fig. 7. Notably, words such as 'also', 'one', 'first', 'people', 'however', 'time', and 'many' appear more than 20,000 times each.[2]

Writer Characteristics: Globally, 1,215 contributors have actively participated in curating handwritten documents, resulting in a diverse dataset encompassing various handwriting styles, camera

[2] Additional word clouds and plots can be found in the supplementary material.

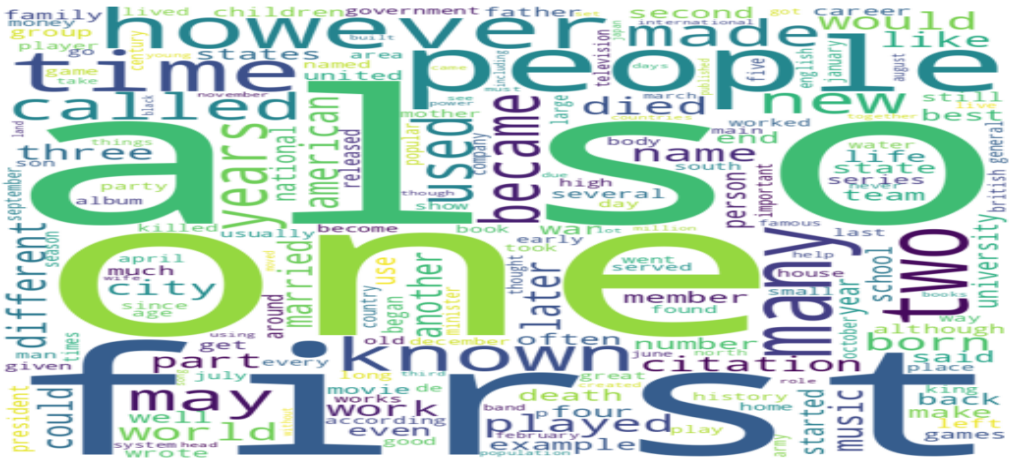


Fig. 6. Visualize the distribution of unique alphabetic words in the dataset through word clouds. It depicts the distribution of all unique alphabetic words in our dataset. The x-axis illustrates each unique word, while the y-axis represents their occurrence on a logarithmic scale.

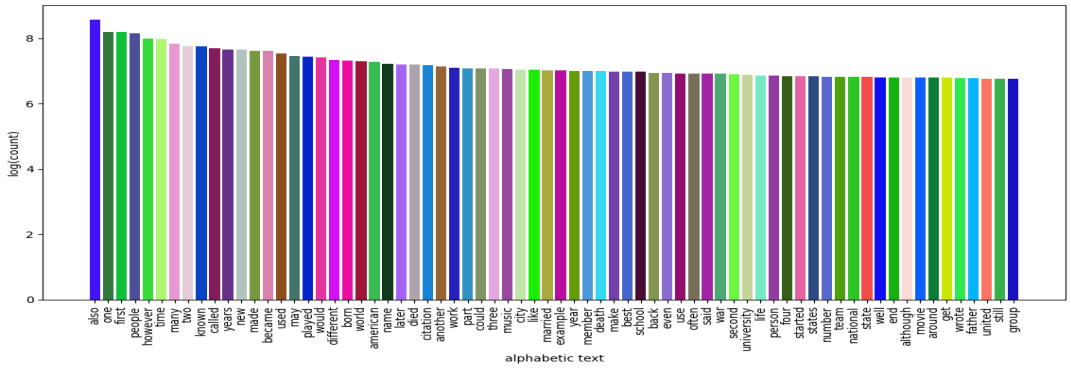
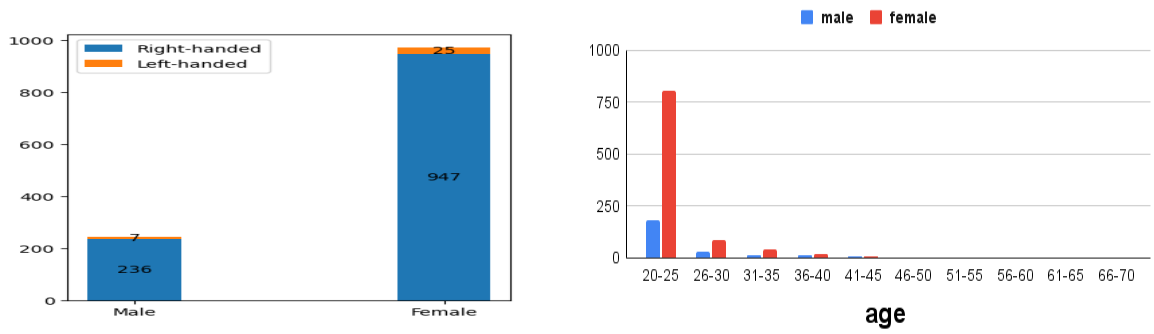


Fig. 7. Presents the frequency distribution of the most common 70 words within the dataset through visualizations. It illustrates the distribution of the top 70 common alphabetic words while plotting.

specifications, scanning methods, and more. The statistical distribution of writers is illustrated in Fig. 8(a), where it is revealed that out of the 1,215 contributors, 972 are female writers, and 243 are male writers. Within the male writers, 7 are identified as left-handed, while 236 are right-handed. Among the female contributors, 25 are left-handed, and the majority, specifically 947, are right-handed. Further demographic details, such as age distribution, are presented in Fig. 8(b). Notably, a significant portion of the contributors falls within the age range of 20 to 40.

Dataset Splits: To furnish an extensive training dataset for deep learning models, our dataset has been partitioned into 521,298 word level images for training, 66,566 word level images for validation, and 169,966 word level images for testing. It will ensure that the models are trained on a large and diverse dataset and will help to improve their accuracy and performance.



(a) (b)

Fig. 8. The statistics provide insights into the demographics of writers collecting handwritten documents. Sub-figure (a) presents data on the distribution of left and right-handed writers among males and females. Sub-figure (b) showcases the demographic distribution of writers categorized by age groups.

Comparison with Existing Datasets: Table 1 comprehensively compares our dataset and existing offline English handwritten text recognition datasets, emphasizing significant distinctions and advantages. Factors such as dataset size, diversity in handwriting styles, and the inclusion of diverse texts are carefully evaluated. Compared to current datasets, ours is notably 13 times and 30 times larger than IAM and GNHK concerning page images, leading to a more extensive collection of unique texts in the dataset. Additionally, our dataset features twice the number of writers compared to IAM, introducing greater diversity in handwriting styles. The experimental section further delves into the potential impact of these differences on the performance and generalization capabilities of models trained on each dataset. This comparative analysis aims to offer a comprehensive understanding of our proposed dataset's distinctive contributions and characteristics within the broader context of existing resources.

4 Benchmark Experiments

4.1 Experimental Settings

Baselines: We adopt the network architecture introduced by Gongidi *et al.* [5] shown in Fig. 9, as a baseline for this experiment. The employed network comprises four key modules: the Transformation Network (TN), Feature Extractor (FE), Sequence Modeling (SM), and Predictive Modeling (PM). The Transformation Network is composed of six plain convolutional layers with 16, 32, 64, 128, 128, and 128 channels. Each layer follows a filter size, stride, and padding size of 3, 1, and 1, respectively, and is succeeded by a 2 $\times$2 max-pooling layer with a stride of 2. The Feature Extractor module adopts the ResNet architecture. The Sequence Modeling incorporates a 2-layer Bidirectional LSTM (BLSTM) architecture with 256 hidden neurons in each layer. Finally, Predictive Modeling employs Connectionist Temporal Classification (CTC) to decode and recognize characters by aligning the feature sequence with the target character sequence. Interested readers can find more details in [5].

Implementation Details: The baseline model undergoes training on a single NVIDIA GeForce GTX 1080 Ti GPU. Input word level images are resized to dimensions of 96 $\times$ 256. Stochastic

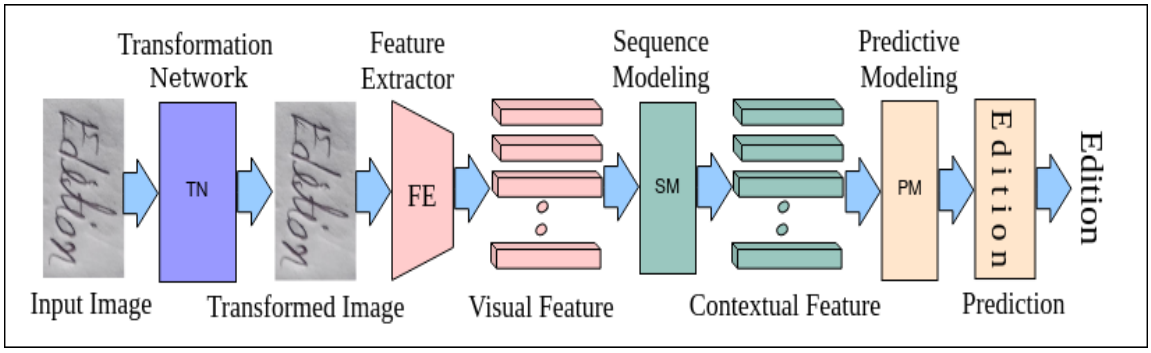


Fig. 9. Processing text recognition through the baseline pipeline.

Gradient Descent (SGD) employs the Adadelta optimizer for all experiments, with a learning rate set to 0.001, a batch size of 64, and momentum fixed at 0.09.

Training/Testing Details: The baseline model undergoes training on the training set of our dataset. Subsequently, we assess the performance of the baseline model on the test sets of our, IAM, and GNHK 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.

4.2 Benchmark Results on Word Level Text Recognition

The performance evaluation outcomes of our baseline model on various offline English handwritten text recognition datasets are displayed in Table 3. The table indicates that the model performs best when trained and tested on the same dataset. More precisely, when the model is trained with IAM and subsequently tested on IAM, trained with GNHK and tested on GNHK, or trained with our dataset and tested on our dataset, it consistently exhibits superior performance within the specific dataset (see 1st row of IAM, GNHK, our dataset in Table 3). This consistent pattern underscores the model's effectiveness when deployed within the dataset on which it was initially trained.

The adaptability and complexity of datasets significantly influence the training and evaluation of diverse algorithms [16]. To scrutinize these aspects, we employ the cross-dataset analysis

Train on	Finetune on	Test Dataset					
		IAM		GHNK		Our	
		CRR	WRR	CRR	WRR	CRR	WRR
IAM	-	89.36	74.76	61.45	32.52	83.09	58.19
	GNHK	82.54	59.61	81.28	60.64	91.17	72.35
	Our	80.77	53.35	75.02	47.82	96.27	86.65
GHNK	-	75.64	49.23	77.22	53.83	87.86	65.64
	IAM	89.19	74.93	65.90	38.41	82.92	58.90
	Our	74.53	45.04	70.72	41.78	95.90	85.49
Our	-	77.16	50.34	70.72	41.97	**96.33**	**86.83**
	IAM	**90.85**	**77.92**	67.59	39.27	89.37	67.84
	GNHK	83.26	60.35	**83.44**	**64.31**	93.89	79.33
Our+IAM+GNHK	-	88.35	71.77	79.57	55.52	96.20	86.33

Table 3. Quantitative results on different English handwritten text recognition datasets. While model pre-trained with our dataset and fine-tuned with respective datasets, the model consistently achieves optimal performance on those respective datasets. The best results are highlighted in bold text.

method [14], wherein a model is trained on one dataset and tested on others. We choose three prominent datasets for our study: IAM, GNHK, and our newly introduced dataset. This systematic approach allows us to assess the generalization capabilities and challenges different datasets pose on algorithmic performance. From the entries in the table (refer to the 1st row corresponding to IAM, GNHK, and our dataset in Table 3), it is evident that when the model is trained with IAM and evaluated on GNHK and our datasets (case-I), it yields the lowest performances on GNHK (61.45 CRR and 32.52 WRR) and our dataset (83.09 CRR and 58.19 WRR). However, in the scenario where the model is trained with GNHK and assessed on IAM and our datasets (case-II), it demonstrates an improvement, with a 7.45 WRR increase compared to case-I for our dataset. Additionally, when the model is trained with our dataset and tested on IAM and GNHK datasets (case-III), it achieves a 1.25 WRR improvement over case-II on the IAM dataset.

We also conducted an additional experiment wherein a model was trained on one dataset and then fine-tuned using a different, new dataset, followed by testing the fine-tuned model on the new dataset. In Experiment-I, the model underwent training with IAM, underwent fine-tuning using GNHK, and was then tested on GNHK. Similarly, in Experiment-II, the model was trained with IAM, fine-tuned on our dataset, and tested on our dataset. Incorporating fine-tuning in both cases resulted in performance improvements, notably achieving higher word recognition rates (28.12 WRR and 30.56 WRR) for both GNHK and our datasets, surpassing the outcomes of the base experiment (case-I). In Experiment-III, the baseline model underwent training with GNHK, was fine-tuned using IAM, and tested on IAM. Simultaneously, in Experiment-IV, the model was trained with GNHK, fine-tuned with our dataset, and then tested on our dataset. In both scenarios, the introduction of fine-tuning led to significant performance improvements (25.7 WRR and 19.85 WRR) for both IAM and our datasets, outperforming the results of the base experiment (case-II). Moving to Experiment-V, the baseline model was initially trained with our dataset, underwent fine-tuning using IAM, and was subsequently tested on IAM. Conversely, in Experiment-VI, the model underwent training with our dataset, fine-tuned with GNHK, and was then tested on GNHK. In both cases, the implementation of fine-tuning resulted in substantial performance enhancements

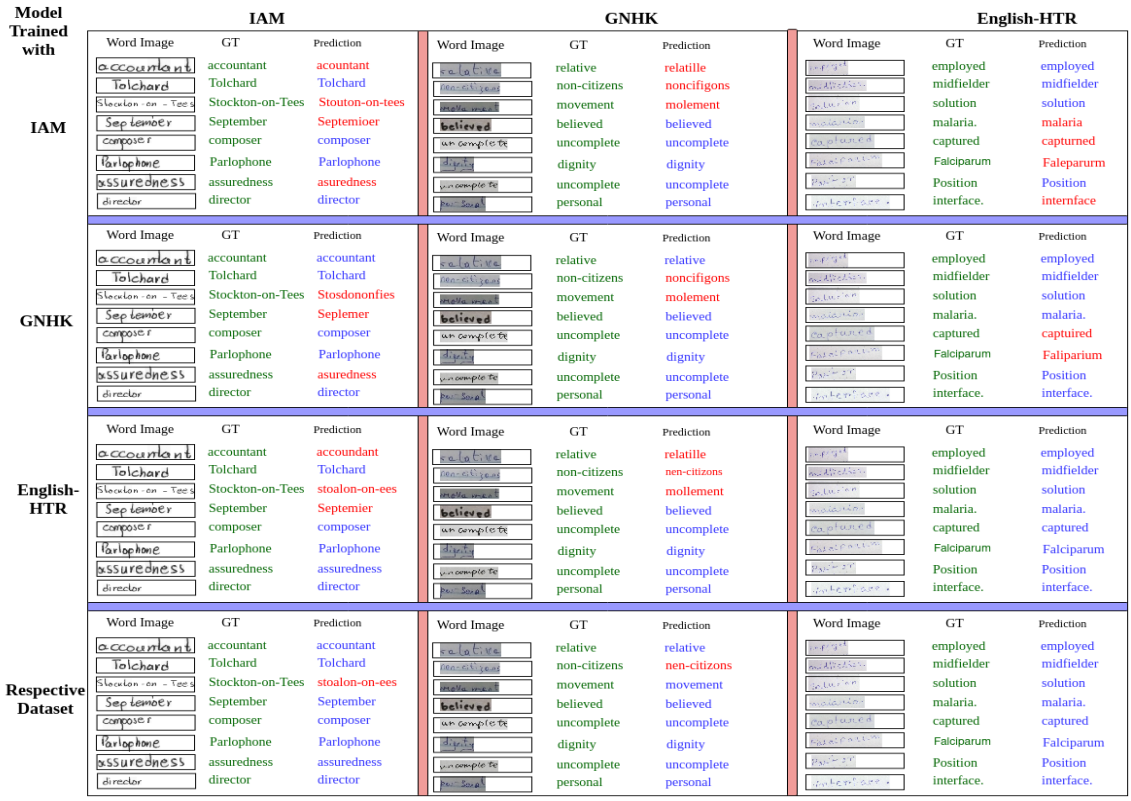


Fig. 10. Displays visual results using various methods. Optimal viewing experience is achieved in color and when zoomed in. The first row showcases results obtained when the model is trained with the IAM dataset. The second row presents results obtained from training the model with the GNHK dataset. The third row illustrates results acquired when the model is trained with our dataset. The fourth row exhibits results obtained when the model is trained with the our dataset and fine-tuned with the respective datasets.

(27.58 WRR and 22.34 WRR) for both IAM and GNHK datasets, surpassing the outcomes of the initial experiment (case-III).

From these experiments, several noteworthy findings emerge: (i) when the model is trained with our dataset and tested on the same dataset (case-III), (ii) when the model is trained with our dataset, fine-tuned on IAM, and subsequently tested on IAM, and (iii) when the model is trained with our dataset, fine-tuned on GNHK, and tested on GNHK, the model consistently achieves the best results, as highlighted by the bold values in Table 3. These experiments underscore the superior generalization capabilities of our dataset compared to the other two existing datasets. Furthermore, they suggest that our dataset can serve as an effective pre-training source to enhance performance on their respective datasets.

Moreover, an additional experiment was conducted, where the model underwent training using all datasets, including IAM, GNHK, and our datasets. Subsequently, evaluations were performed on these datasets, as illustrated in the last row of Table 3. The results from the table indicate that the performance of combined training is lower than that of pre-training and fine-tuning, as evidenced

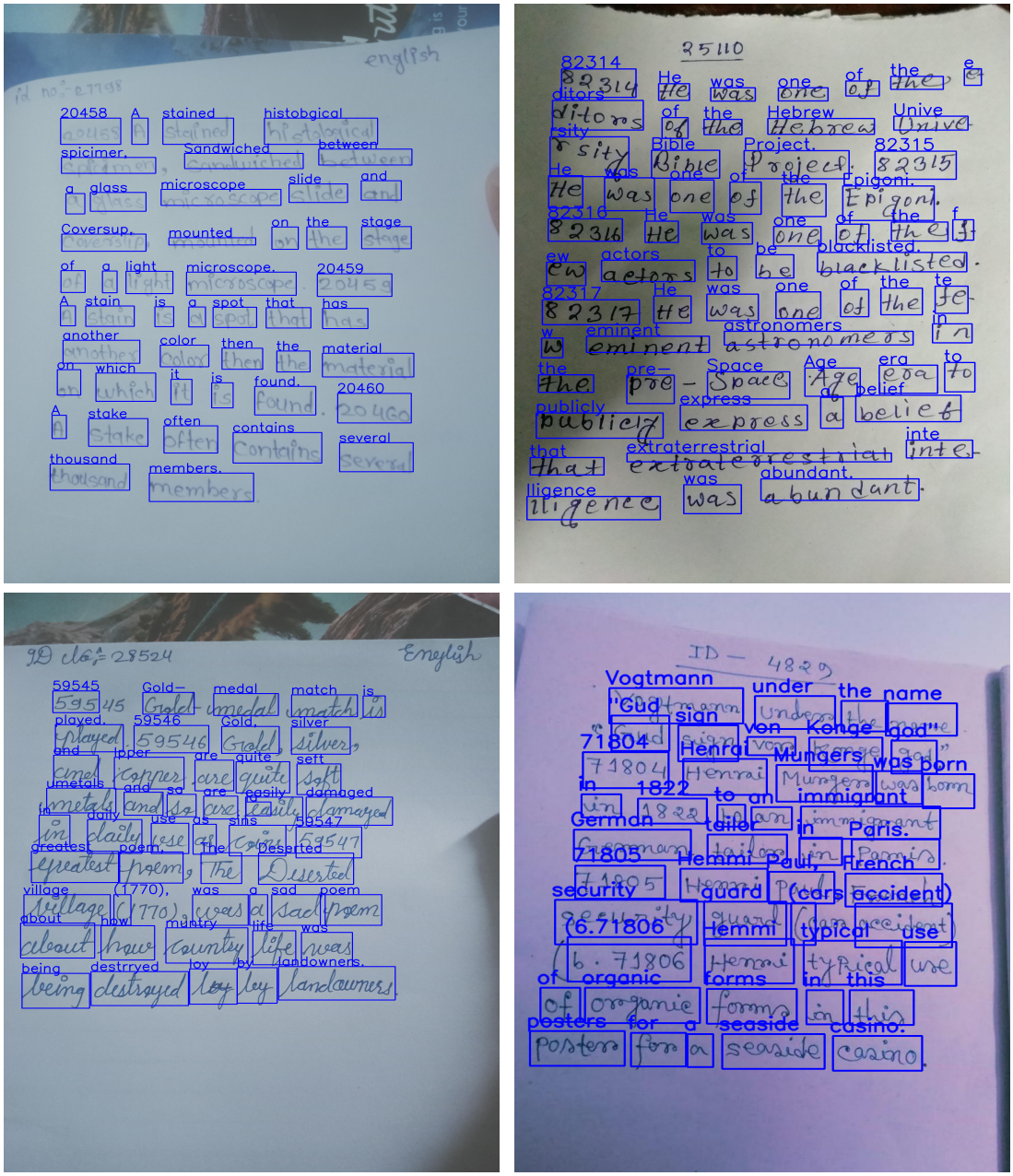


Fig. 11. Examples of page level results for word detection and recognition. Blue bounding boxes and texts highlight the results of word detection and recognition.

by the lower word recognition rates in the case of IAM (6.15 WRR), GNHK (8.79 WRR), and our

dataset (0.5 WRR). It suggests that training with a combined dataset is less effective than the pre-training and fine-tuning approach, as demonstrated in Experiment-I through Experiment-VI.

In Fig. 10, visual results are presented, showcasing the model's performance when trained with different datasets. Notably, when the model is trained solely with the IAM dataset, it inaccurately predicts four out of eight words. However, incorporating fine-tuning with IAM after training with our dataset reduces this error to just one incorrectly predicted word out of eight on IAM. Similarly, training the model with the GNHK dataset leads to two erroneous predictions out of eight words in the GNHK dataset. Conversely, when the model is initially trained with our dataset and subsequently fine-tuned with GNHK, the number of incorrect predictions is reduced to just one out of eight words in the GNHK dataset. The visual results underscore the effectiveness of pre-training with our dataset in minimizing incorrect predictions across different datasets. This observation suggests that a pre-training strategy with a broader dataset can significantly enhance the model's overall predictive accuracy.

4.3 Page Level Results

We provide a few page level results in Fig. 11 for visual illustration. The results depict both word detection and recognition in blue colored bounding boxes and blue colored text. The results of page level handwritten text recognition demonstrate the effectiveness of our approach in accurately transcribing handwritten text from entire pages. The results show that we have significantly improved our model's performance. These results underscore the potential of our methodology to handle the complexities and variability inherent in handwritten documents, paving the way for enhanced document digitization and text recognition applications.

5 Conclusions

We introduced *IIIT-HW-English-Word*, a large and diverse offline handwritten text recognition dataset. This dataset comprises unconstrained camera-captured images of English handwritten documents gathered from various regions in India. With a total of 20,800 document pages contributed by 1,215 distinct writers, the dataset provides a rich and varied collection. Among the 20,800 page images, 757,830 instances of words were identified, encompassing alphabetic, numeric, stop-words, and other categories. Of these occurrences, 174,701 are unique. Our paper presents benchmark results on text recognition using well-established architectures. The experiments indicate that training the model with our dataset enhances its performance on established offline handwritten text recognition datasets. Thus, our dataset improves the model's performance across existing datasets. Future research avenues could explore end-to-end approaches, integrating localization and recognition within the same framework. We welcome contributions from researchers and developers to create new models leveraging this dataset.

Acknowledgments

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

References

1. Aksan, E., Pece, F., Hilliges, O.: Deepwriting: Making digital ink editable via deep generative modeling. In: Proceedings of the CHI conference on human factors in computing systems. pp. 1–14 (2018)
2. Bukhari, S.S., Shafait, F., Breuel, T.M.: The iupr dataset of camera-captured document images. In: CBDAR. pp. 164–171 (2012)
3. Fischer, A., Frinken, V., Fornés, A., Bunke, H.: Transcription alignment of latin manuscripts using hidden markov models. In: WHDIP. pp. 29–36 (2011)
4. Fischer, A., Keller, A., Frinken, V., Bunke, H.: Lexicon-free handwritten word spotting using character hmms. Pattern recognition letters **33**(7), 934–942 (2012)
5. Gongidi, S., Jawahar, C.: IIIT-INDIC-HW-Words: A dataset for indic handwritten text recognition. In: ICDAR. pp. 444–459 (2021)
6. Grosicki[1], E., Carre, M., Brodin, J.M., Geoffrois[1], E.: Rimes evaluation campaign for handwritten mail processing (2008)
7. Lee, A.W., Chung, J., Lee, M.: Gnhk: A dataset for english handwriting in the wild. In: ICDAR. pp. 399–412 (2021)
8. Marti, U.V., Bunke, H.: The IAM-database: an english sentence database for offline handwriting recognition. IJDAR **5**, 39–46 (2002)
9. Potanin, M., Dimitrov, D., Shonenkov, A., Bataev, V., Karachev, D., Novopoltsev, M.: Digital peter: Dataset, competition and handwriting recognition methods. arxiv preprint, 2021. Source:¡ https://arxiv.org/abs/2103.09354 (2021)
10. Sánchez, J.A., Romero, V., Toselli, A.H., Vidal, E.: ICFHR2014 competition on handwritten text recognition on transcriptorium datasets (htrts). In: ICFHR. pp. 785–790 (2014)
11. Sanchez, J.A., Romero, V., Toselli, A.H., Vidal, E.: Icfhr2016 competition on handwritten text recognition on the read dataset. In: ICFHR. pp. 630–635 (2016)
12. Tarride, S., Boillet, M., Moufflet, J.F., Kermorvant, C.: SIMARA: a database for key-value information extraction from full pages. arXiv preprint arXiv:2304.13606 (2023)
13. Tarride, S., Faine, T., Boillet, M., Mouchère, H., Kermorvant, C.: Handwritten text recognition from crowdsourced annotations. arXiv preprint arXiv:2306.10878 (2023)
14. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528 (2011)
15. Viard-Gaudin, C., Lallican, P.M., Knerr, S., Binter, P.: The ireste on/off (ironoff) dual handwriting database. In: ICDAR. pp. 455–458 (1999)
16. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 3239–3259 (2021)
17. Zimmermann, M., Bunke, H.: Automatic segmentation of the IAM off-line database for handwritten english text. In: ICPR. vol. 4, pp. 35–39 (2002)