ICDAR 2024 Competition on Recognition and VQA on Handwritten Documents

Ajoy Mondal[1][0000−0002−4808−8860], Vijay Mahadevan[2][0000−0002−3337−2607], R. Manmatha[2][0000−0003−2315−8583], and C. V. Jawahar[1][0000−0001−6767−7057]

[1] International Institute of Information Technology, Hyderabad, India
[2] AWS AI Labs, San Francisco, California, USA

Abstract. This paper presents a competition report on the Recognition and Visual Question Answer on Handwritten Documents towards deeper understanding of handwritten multilingual documents (ICDAR 2024-HWD) held at the 18th International Conference on Document Analysis and Recognition (ICDAR 2024). Documents are in English or Indian Languages. Earlier editions related to recognition of Indian handwriting were held in conjunction with ICFHR 2022 and ICDAR 2023. A related DocVQA task was held in DAS 2020. This edition proposes three main tasks: *Isolated Word Recognition (Task A)*, *Page Level Recognition and Reading (Task B)*, and *Visual Question Answers on Handwritten Documents (Task A)*. While Task A was already part of our previous competitions, we bring in new data as part of this edition. Task B and Task C are novel additions for this year. By attracting researchers with experience in printed and handwritten documents, we aim to establish benchmarks that significantly contribute to the academic literature in this field. A total of thirty-two teams around the world registered for this competition. Among them, only ten teams submitted their results along with algorithm details. The winning team, **TSNUK**, achieved an average 98.00% Character Recognition Rate (CRR) and 94.26% Word Recognition Rate (WRR) across four languages for Task A: Isolated Word Recognition. **IndependentOCR** excelled in Task B: Page Level Recognition and Reading, with 76.32% average Page Level Character Recognition Rate (PCRR) and 62.57% average Page Level Word Recognition Rate. The team **PA_VCG** won Task C: Visual Question Answering on Handwritten Documents with a 0.643 ANLS score.

Keywords: Handwritten documents, word detection, word recognition, reading order, recognition of paragraph, and visual question answer.

1 Introduction

Handwritten OCRs are now becoming practical with many commercial APIs, solutions and use cases. However, they are still popular only for English or Latin scripts (barring exceptions in some limited Asian scripts). We need data sets and knowledge sharing for extending these results to other languages like Indian languages. Even for English, the publicly available data is limited, making academic

research lagging behind the industrial solutions. Even when a full fledged recognition/understanding is difficult, today's document understanding systems can meet the user need with Question Answering (QA) tasks. They were also defined for the Document Question Answering (DocVQA). This may be the time to look into information extraction from handwritten documents where the recognition, layout (structure and content) is still challenging.

Handwritten text recognition poses unique challenges due to several factors. (i) Significant Style Variability - handwriting exhibits considerable style variability, making it a complex task to develop robust recognition algorithms that accommodate diverse writing forms. (ii) Content Variability - handwritten content spans a wide spectrum, ranging from formal text to informal notes. This variability demands adaptability in recognition models to interpret diverse content types effectively. (iii) Temporal Changes - handwriting may evolve, introducing an additional layer of complexity. Adapting to changes in an individual's writing style necessitates continuous refinement of recognition models. The inherent challenges of handwritten text recognition serve as a potent motivator for researchers, sparking interest and driving exploration in this demanding and dynamic field. In essence, OCR bridges the visual and machine-readable realms, explicitly focusing on the complexities of handwritten text recognition. The ongoing pursuit of solutions in this interdisciplinary and challenging domain reflects the resilience of researchers in pushing the boundaries of what OCR can achieve.

While handwritten text recognition has made significant strides for certain languages such as English [8, 22, 12], Chinese [26, 25, 21], Arabic [15, 9], and Japanese [14, 19], a considerable gap persists for many languages globally. Unfortunately, several Indian scripts and languages are underrepresented in OCR research efforts, placing them at risk of being left behind in the technological landscape. Only a handful of the 22 languages spoken in India have received attention, primarily for communication purposes. The pressing need for research on text recognition for Indic scripts and languages cannot be overstated. Languages such as Hindi, Bengali, and Telugu, among the most spoken in India [11], urgently need OCR solutions tailored to their unique characteristics. Indic scripts pose specific challenges, making handwritten text recognition more demanding than Latin scripts. In most Indic scripts, forming conjunct characters, where two or more characters combine, is a common feature [2]. This complexity introduces intricacies not present in scripts like English. Compared to the relatively straightforward 52 unique characters (upper and lower case) in English, most Indic scripts boast over 100 unique basic Unicode characters [20]. This richness in character sets demands specialized attention in OCR systems to ensure accurate recognition.

In the previous editions of the competitions, ICFHR 2022 IHTR and ICDAR 2023 IHTR [18], we provided an existing training set, an existing test set, and a newly created test set. In ICFHR 2022 IHTR, eleven participants registered for the competition, and five teams submitted results. At ICDAR 2023 IHTR, eighteen teams registered for the competition, and eight presented results. Multiple participants used the newer methods based on state-of-the-art

architectures. The proposed competition continues this effort with even more unique datasets and introduces two novel tasks: page level recognition and reading, and visual question answers on handwritten documents. Our challenge, hosted on `https://ilocr.iiit.ac.in/icdar_2024_hwd/`, is centered around handwritten document recognition. This competition stands as a dynamic catalyst, igniting the passion and creativity of researchers to pioneer groundbreaking solutions in the realm of handwritten document analysis. By providing a platform for innovation and algorithm design, it serves as a driving force, inspiring participants to push the boundaries of what is achievable in understanding and interpreting handwritten documents.

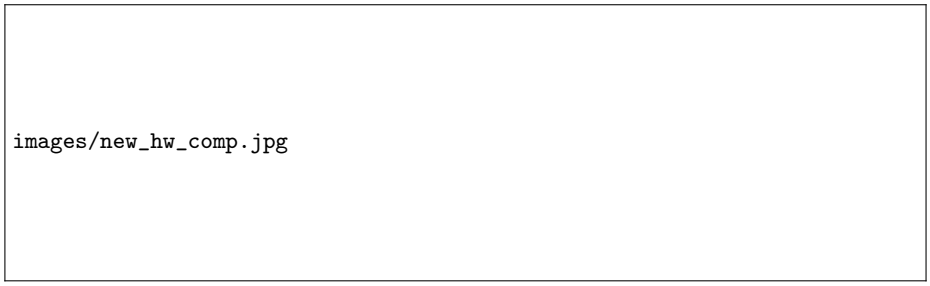

Fig. 1. Proposed Competition Tasks.

2 Specific Challenge Tasks

For the challenge, we focus on the following languages:

– Bengali, English, Hindi, Telugu

We propose following three main tasks (see Fig. 1) on handwritten documents.

1. Task A: Isolated Word Recognition
2. Task B: Page level Recognition and Reading
3. Task C: Visual Question Answers on Handwritten Documents

Task A: Isolated Word Recognition — Task A, a consistent feature from our prior competitions, including ICFHR 2022 IHTR and ICDAR 2023 IHTR, focuses on designing robust algorithms for isolated word recognition. The primary objective is to create algorithms that accurately recognize each word within a provided set of word images.

Task B: Page Level Recognition and Reading — Task B is an innovative addition to this year's competition, offering participants a fresh challenge in handwritten text recognition. The primary objective is to develop a robust

algorithm capable of recognizing complete pages and preserving the reading order. Participants have the flexibility to approach the task with word or line-level segmentation, employing techniques that enhance recognition through finer granularity. The competition welcomes submissions for those opting not to use segmentation (word or line-level). If the performance without segmentation is comparable to the best scores achieved with segmentation, such submissions will be recognized as exceptional.

Task C: Visual Question Answer on Handwritten Documents — Beyond mere recognition, the contemporary focus in document analysis extends to extracting meaningful information tailored to users' needs. Drawing inspiration from pioneering works such as DocVQA [16] and related tasks [24, 23, 27] for printed documents, this challenge introduces a VQA task specifically dedicated to handwritten documents. Based on the practices in the past work (such as DocVQA [16]), we use Average Normalized Levenshtein Similarity (ANLS) as the evaluation metric. These metrics provide a robust and comprehensive assessment of the performance of VQA algorithms on handwritten documents. The purpose of this task is multi-fold. It catalyzes advancing research in the extraction of information from handwritten documents. Participants are encouraged to explore innovative approaches beyond recognition and delve into deeper layers of document understanding. A fundamental question posed by this task is the transfer of knowledge - from printed to handwritten documents or between different classes of documents. This inquiry opens avenues for exploring the adaptability and transferability of algorithms across diverse document types.

3 Dataset

We collect handwritten pages of four languages, English, Hindi, Bengali, and Telugu, from native writers all over India. We created a web-based online data collection tool for collecting handwritten pages corresponding to the given text paragraphs from native writers. There is no restriction while writing. After writing the page, the writers scan or capture it by mobile and upload handwritten page images into the tool. We collect near to 4K handwritten page images written by 100-200 writers per language. Due to the unconstrained writing, several complexities like significant skew, non-uniform gap between words and lines, and overlapping words. The mobile camera also introduces several other complex issues like orientation, blurring, extra noisy background, cutting of boundary lines, and reflection while capturing pages. A few sample images are shown in Fig. 2. We annotate all images manually. For a handwritten page, the ground truth contains bounding boxes, reading order, and text transcription of words in the page. The dataset is divided into a training set of near to 3K and a test set of near to 1K page images per language. We provide ground truths only for the training set. Table 1 shows statistics of dataset used for this competition.

Fig. 2. Shows few sample images of the dataset. First Row: English handwritten pages, Second Row: Bengali handwritten pages, Third Row: Hindi handwritten pages, Fourth Row: Telugu handwritten pages.

IHWWR-1.0 Dataset for Task A		
Language	**Training Set**	**Test Set**
Bengali	79663	17108
English	85020	13601
Hindi	85585	20511
Telugu	77625	9585
PLHWTR-1.0 Dataset for Task B		
Language	**Training Set**	**Test Set**
Bengali	3108	700
English	3500	407
Hindi	3500	868
Telugu	3496	700
SP-HWVQA-1.0 Dataset for Task C		
Language	**Training Set**	**Test Set**
	#Image, #question	**#Image, #question**
English	1000, 4000	1000, 1000

Table 1. Shows the statistics of the used dataset in this competition.

4 Evaluation Metric

Task A: Isolated Word Recognition: We use two famous evaluation metrics, 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 submitted word level recognizers. Error Rate (ER) is defined as

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

where S indicates the number of substitutions, D indicates the number of deletions, I indicates the number of insertions, and N number of instances in reference text. In the case of CER, Eq. 1 operates on character level, and in the case of WER, Eq. 1 operates on word level. Recognition Rate (RR) is defined as

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

In the case of CRR, Eq. 2 operates on character level, and in the case of WRR, Eq. 2 works on word level.

Task B: Page Level Recognition and Reading: We average CRR and WRR, defined in Eq. 2, over all pages in the test set and define new measure average page level CRR (PCRR) and average page level WRR (PWRR).

$$PCRR = \frac{1}{L}\sum_{i=1}^{L} CRR(i), \tag{3}$$

and

$$PWRR = \frac{1}{L}\sum_{i=1}^{L} WRR(i), \tag{4}$$

where $CRR(i)$ and $WRR(i)$ are CRR and WRR of i^{th} page in the test set.

Task C: Visual Question Answers on Handwritten Documents: Similar to the past works [16, 24], we use Average Normalized Levenshtein Similarity (ANLS) as the evaluation metric for this task. ANLS is given by Eq. 5, where N is the total number of questions, M are possible ground truth answers per question, $i = 0...N$, $j = 0...M$ and o_{q_i} is the answer to the i^{th} question q_i.

$$\begin{aligned} ANLS &= \frac{1}{N}\sum_{i=0}^{N}\Big(\max_{j}\ s(a_{ij}, o_{q_i})\Big) \\ s(a_{ij}, o_{q_i}) &= \begin{cases} 1 - NL(a_{ij}, o_{q_i}), & \text{if } NL(a_{ij}, o_{q_i}) < \tau. \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{5}$$

where $NL(a_{ij},\ o_{q_i})$ is the normalized Levenshtein distance (ranges between 0 and 1) between the strings a_{ij} and o_{qi}. The value of τ can be set to add softness toward recognition errors. If the normalized edit distance exceeds τ, it is assumed that the error is because of an incorrectly located answer rather than an OCR mistake.

5 Methods

Thirty-two participants around the world registered for the competition. However, we got submissions from ten of them. These ten teams are (i) TSNUK, Taras Shevchenko National University of Kyiv, Ukraine, (ii) VitaminC_PA, Ping An Property & Casualty Insurance Company of ChinaLtd, China, (iii) Team-LTU, EISLAB, SRT, Luleå University of Technology, (iv) PA_VCG, Ping An Property & Casualty Insurance Company of China, Ltd, China, (v) IndependentOCR, NIT Surat, India, (vi) IndicOCR Group, Digital University Kerala, India, (vii) TeamOCR, IIIT Nagpur, India, (viii) OCRIndia, India, (ix) IndicOCR, India and (x) GroupIndia, India.

5.1 Methods for Task A

VitaminC_PA: — The team chose PARSeq (Permuted Autoregressive Sequence Models) [5] as base model, which has strong ability to integrate context-free non-AR and context-aware AR inference, along with iterative refinement using bidirectional context. It is optimal in accuracy versus parameter count, FLOPS, and latency due to its unified structure and parallel token processing mechanism. To ensure a higher performance when generalize the base model to

our target tasks, the team pre-trains the neural network on large-scale real-world datasets including COCO-Text (COCO) , RCTW17, Uber-Text (Uber), ArT , LSVT , MLT19 , TextOCR and ReCTS , as well as synthetic datasets MJSynth (MJ) and SynthText (ST). During the pre-training process of the PARSeq model, several data preprocessing steps: tokenization, position embedding, label processing, case sensitivity handling, character filtering and data augmentation, are performed to generate training data. The team finetunes the model on the Task A training datasets using a Nvidia 3090 GPU with the batch size 384. Data augmentation methods are applied, including RandAugment,GaussianBlur and PoissonNoise. We adopt an AdamW optimizer with a base learning rate 0.0007. OneCycleLR method is used to dynamically adjust the learning rate.

IndependentOCR: This team utilized PARSeq [5] for their experiments. They pre-trained the models using printed word-level images in English, Hindi, Bengali, and Telugu. Subsequently, they fine-tuned the model with the respective required training sets.

TeamOCR: This team employs the CRNN architecture [7], consisting of four main modules: the Transformation Network (TN), Feature Extractor (FE), Sequence Modeling (SM), and Predictive Modeling (PM). The Transformation Network includes six plain convolutional layers, each followed by a max-pooling layer with a size of 2×2 and a stride of 2. The Feature Extractor module uses the ResNet architecture, while the Sequence Modeling module comprises a 2-layer Bidirectional LSTM (BLSTM) with 256 hidden neurons per layer. Finally, the Predictive Modeling module utilizes Connectionist Temporal Classification (CTC) for character decoding and recognition, aligning the feature sequence with the target character sequence.

IndicOCR Group: This team also used CRNN architecture for Task A. They created separate model corresponding four languages English, Hindi, Bengali, and Telugu. They pre-trained each model with printed word level images and then fine-tuned with respective required training sets.

5.2 Methods for Task B

IndependentOCR: The method follows two steps - (i) detect individual words in the page using CRAFT [3], and (ii) finally each detected word is recognized by PARSeq [5].

TeamOCR: The method follows two steps - (i) detect individual words in the page using DocTR [17], and (ii) then each detected word is recognized by CRNN [7].

IndicOCR Group: This team used open source easyOCR [1] for page level recognition.

5.3 Methods for Task c

PA_ VCG: The solution utilizes the open-source InternVL as the foundation model and they further perform LoRA finetuning on it using the OCR results predicted by QwenVL-max model for supervision. Then we design thinking chains for different types of questions and conduct multi-round conversations to produce final answers. The proposed method contains two steps, i.e., finetuning the InternVL model and multi-round QA (questioning and answering). (i) **Finetuning InternVL Model:** The team chooses the recent open-source large vision language model (LVLM) InternVL [6] as the foundation model since it exhibits good potentials in handwritten OCR tests. Despite such advantages, the OCR capability of InternVL is still not as good as that of the closed-source QwenVL-max model [4]. Consequently, the team uses the QwenVL-max API to generate structured OCR results (in json format) of the handwritten census records. Then the team performs LoRA finetuning on InternVL model using the generated OCR results for supervision. (ii) **Multi-round QA** the team conducts three rounds of dialogues to predict final answers. (1) Generating OCR results. The finetuned InternVL model isrequired to predict the structured OCR results of images. (2) Asking questions. Based on the OCR results, the team asks the LVLM to answer the question and explain why. In the experiments, requesting explanations can further improve the precision of answers. Additionally, for complex inquiries that necessitate a deeper level of reasoning and statistical analysis, the team designs thinking chains to guide the model towards more accurate conclusions. (3) Formatting. We make a statistics on the questions from training set, and categorize them into 25 classes. The team then predicts the question type of each testing question and randomly construct QA examples accordingly for providing the guidance in answer formatting.

Team_ EISLAB: The team used SPHINX [13] for this task. teh SPHINX is fine-tuned on the handwritten question and answering pairs from the required training set.

IndependentOCR: This team used pre-trained PARSeq [5] to recognize the the handwritten text and then these text are used trained BERT [10] model for answering the questions.

6 Competition Results

6.1 Results Analysis for Task A

Quantitative Results: — Table 2 and Table 3 show the teams' results for individual languages. The team VitaminC_PA obtained the best CRR (98.83%) and WRR (96.12%) scores for English. At the same time, TSNUK obtained the second-best result (CRR 98.28% and WRR 94.68%) in English. The team TSNUK obtained the best performance (CRR 97.55%, 99.06% and 96.72%) and (WRR 94.37%, 97.82%, and 90.21%) for Bengali, Hindi and Telugu languages,

Bengali			**English**		
Team	**CRR**	**WRR**	**Team**	**CRR**	**WRR**
VitaminC_PA	91.77	79.80	VitaminC_PA	**98.83**	**96.12**
TSNUK	**97.55**	**94.37**	TSNUK	98.28	94.68
IndependentOCR	82.78	72.23	IndependentOCR	88.28	83.94
TeamOCR	78.53	63.21	TeamOCR	87.38	81.19
IndicOCR Group	75.28	60.73	IndicOCR Group	85.88	81.03

Table 2. Shows CRR and WRR of Bengali and English languages for Task A of different teams. The bold value indicates the best results.

Hindi			**Telugu**		
Team	**CRR**	**WRR**	**Team**	**CRR**	**WRR**
TSNUK	**99.06**	**97.82**	TSNUK	**96.72**	**90.21**
IndependentOCR	90.53	86.21	VitaminC_PA	94.62	80.11
TeamOCR	88.25	85.63	IndependentOCR	87.39	75.12
IndicOCR Group	86.82	83.10	TeamOCR	85.92	75.03
VitaminC_PA	70.90	21.73	IndicOCR Group	83.22	73.82

Table 3. Shows CRR and WRR of Hindi and Telugu languages for Task A of different teams. The bold value indicates the best results.

Team	**CRR**	**WRR**
TSNUK	**98.00**	**94.26**
IndependentOCR	87.24	79.37
TeamOCR	85.02	76.26
IndicOCR Group	82.80	74.67
VitaminC_PA	89.03	69.44

Table 4. Shows average CRR and WRR over four languages for Task A of different teams. The bold value indicates the best results.

respectively. VitaminC_PA obtained the best results in English, and TSNUK obtained the best in Bengali, Hindi and Telugu. However, when we averaged the performance over four languages, Bengali, English, Hindi, and Telugu, the team TSNUK obtained the best performance (average CRR 98.00% and average WRR 94.26%). The team IndependentOCR obtained the second best performance (average CRR 87.24% and WRR 79.37%). VitaminC_PA obtained the least performance (WRR 69.44%). Table 4 shows the average CRR and WRR over four languages for all the participating teams.

Qualitative Results: — Fig. 3 showcases several samples of qualitative results obtained by various teams. For the Bengali language, the TSNUK team achieved correct predictions for all sample word-level images, demonstrating their model's robustness in handling the intricacies of Bengali script. In contrast, other teams made incorrect predictions for two samples, with one error arising due to the pres-

Fig. 3. shows visual results obtained by various teams. Blue colored text indicates ground truths, Red colored text indicates wrongly recognized text.

ence of a conjunct character in the word. In the case of English, both TSNUK and VitaminC_PA teams made only one mistake each, which was due to character ambiguity. Their models are proficient in English text recognition, handling common variations and ambiguities well. However, the other three teams incorrectly recognized two words, indicating potential areas for improvement in their models' handling of ambiguous characters. The TSNUK team again showed superior performance for Hindi by correctly predicting all outputs. Other teams, however, struggled significantly, making four incorrect predictions out of six words. The errors were primarily due to conjugate characters and upper matra, which complicate Hindi text recognition. In the case of Telugu, all teams faced challenges, particularly with words containing upper and lower matra. The figure highlights that the TSNUK team was notably more successful in recognizing words with these features, indicating their model's advanced handling of Telugu script's unique characteristics.

Overall, Fig. 3 illustrates the TSNUK team's consistent ability to predict words across different languages and script complexities correctly. Their model's effective handling of conjugate characters, upper matra, and lower matra underscores their advanced techniques and fine-tuning processes. This performance highlights the importance of addressing language-specific nuances in OCR models, particularly for scripts with complex orthographic rules.

6.2 Results Analysis for Task B

Quantitative Results: — Table 5 and Table 6 present Page Level CRR and WRR for individual languages. The team IndependentOCR obtained the best performance for all four languages. It is because of pre-trained PARSeq with printed word level images as recognizers to recognize the individual word on a page. The team TeamOCR obtained the second-best performance in all languages. Table 5 shows average page level CRR and WRR for all teams over

Bengali			**English**		
Team	**PCRR**	**PWRR**	**Team**	**PCRR**	**PWRR**
IndependentOCR	**75.39**	61.23	IndependentOCR	78.45	66.26
TeamOCR	71.25	58.38	TeamOCR	74.89	63.21
IndicOCR Group	68.45	52.82	IndicOCR Group	71.34	59.05
OCRIndia	65.23	46.72	OCRIndia	69.56	56.27
IndicOCR	62.38	43.56	IndicOCR	65.81	54.92
GroupIndia	58.50	41.28	GroupIndia	63.03	53.80

Table 5. Shows PCRR and PWRR of Bengali and English languages for Task B of different teams. The bold value indicates the best results.

Hindi			**Telugu**		
Team	**CRR**	**WRR**	**Team**	**CRR**	**WRR**
IndependentOCR	**77.21**	63.43	IndependentOCR	**74.23**	**59.36**
TeamOCR	75.47	59.54	TeamOCR	72.51	57.68
IndicOCR Group	74.90	54.53	IndicOCR Group	70.72	53.43
OCRIndia	71.56	**66.28**	OCRIndia	67.36	56.26
IndicOCR	68.78	58.45	IndicOCR	64.82	56.19
GroupIndia	65.23	56.76	GroupIndia	62.24	55.60

Table 6. Shows PCRR and PWRR of Hindi and Telugu languages for Task B of different teams. The bold value indicates the best results.

Team	**Average PCRR**	**Average PWRR**
IndependentOCR	**76.32**	**62.57**
TeamOCR	73.53	59.70
IndicOCR Group	71.35	54.95
OCRIndia	68.42	56.38
IndicOCR	65.44	53.28
GroupIndia	62.25	51.86

Table 7. Shows average PCRR and PWRR over four languages for Task B of different teams. The bold value indicates the best results.

all languages. The team IndependentOCR obtained the best average page level CRR (76.32%) and WRR (62.57%) over all languages.

Qualitative Results: — Fig. 4 showcases several samples of qualitative results obtained by various teams. The IndependentOCR team predominantly produced correct predictions for the English language, although they occasionally failed to recognize words starting with capital letters. It indicates a minor limitation in their model's ability to handle capitalization, which is crucial for proper noun recognition and sentence structuring. Other teams, however, exhibited a higher frequency of errors in their predictions, underscoring the superior performance of the IndependentOCR team in English text recognition. Their model's robust-

ness in handling typical English text complexities was evident. Regarding Indic languages, the errors were more pronounced across all teams, including IndependentOCR. This increase in errors can be attributed to the inherent complexity of Indic scripts, which often involve intricate characters and diacritical marks. The language complexity poses significant challenges for OCR models, leading to a higher error rate.

Despite these challenges, Fig. 4 illustrates that the IndependentOCR team achieved the best results on the English and Indic language sample images. Their model's relative success in handling the diverse and complex nature of the sample texts demonstrates their advanced techniques and fine-tuning processes. This achievement highlights the potential for further improvements in OCR technology, particularly for non-Latin scripts, where language-specific nuances need to be more effectively addressed.

Fig. 4. shows visual results obtained by various teams. Blue colored text indicates ground truths, Red colored text indicates wrongly recognized text.

6.3 Results Analysis for Task C

Team	**ANLS**
PA_VCG	**0.643**
Team_LTU	0.385
IndependentOCR	0.235
TeamOCR	0.138
IndicOCR Group	0.113

Table 8. Shows ANLS for Task C of different teams. The bold value indicates the best results.

Table 8 presents the quantitative results of all teams for Task C. The team PA_VCG achieved the highest score with an ANLS of 0.643. This success can be attributed to their two-step fine-tuning of the InternVL model and the implementation of multi-round QA (questioning and answering) for text recognition and answering. The effectiveness of their two-step process was key to their top performance. In contrast, Team_LTU secured the second position with an ANLS score of 0.385. The IndicOCR Group had the lowest performance, with an ANLS score of 0.113.

Qualitative Results: — Fig. 5 presents a selection of qualitative results for Task C. The figure demonstrates that PA_VCG and Team_LTU successfully provided correct answers for the selected samples. It indicates their models' robustness and effectiveness in handling the specific challenges posed by Task C, which likely involves complex text recognition and understanding tasks. The success of PA_VCG and Team_LTU can be attributed to their advanced techniques and fine-tuning processes. PA_VCG, for instance, utilized a two-step fine-tuning approach with the InternVL model and incorporated multi-round QA (questioning and answering) strategies. This meticulous process allowed them to achieve high accuracy and reliability in their predictions. On the other hand, the IndependentOCR team struggled with the selected samples and failed to predict the correct answers. This discrepancy suggests that their model may need further refinement to handle the complexities of Task C effectively. The challenges faced by IndependentOCR could be due to various factors, such as insufficient training on diverse datasets, less effective fine-tuning methods, or limitations in their model architecture.

The figure underscores the importance of robust training techniques and comprehensive model evaluation. Teams like PA_VCG and Team_LTU that invest in thorough fine-tuning and innovative approaches tend to perform better in complex OCR tasks. It also highlights areas for improvement for other teams, emphasizing the need for continuous development and adaptation to enhance model performance. In summary, Fig. 5 showcases the varying levels of success among the teams and illustrates the critical role of advanced methodologies in achieving accurate and reliable OCR results. The performance of PA_VCG and Team_LTU sets a benchmark for others, while the challenges faced by IndependentOCR provide valuable insights for future enhancements.

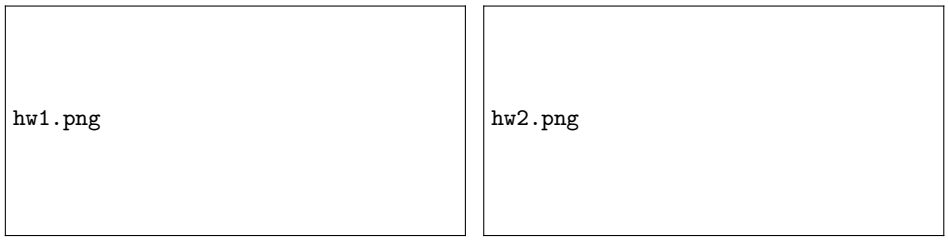

Fig. 5. Shows a few sample outputs of all teams.

7 Conclusion

This competition is a significant motivator for researchers with prior experience in handwritten OCR, encouraging them to establish benchmarks and contribute to the academic literature in this field. Despite thirty-two teams registering, only ten submitted their results and algorithm details, reflecting the challenge's rigorous standards and complexity. In Task A, the team TSNUK emerged victorious, achieving an impressive average Character Recognition Rate (CRR) of 98.00% and Word Recognition Rate (WRR) of 94.26%. This high level of accuracy underscores their model's proficiency in recognizing handwritten characters and words, setting a new benchmark for future research. For Task B, the IndependentOCR team won by attaining an average page-level CRR (PCRR) of 76.32% and a page-level WRR (PWRR) of 62.57%. Their success highlights the effectiveness of their approach in handling the complexities of page-level text recognition, demonstrating their model's robustness and reliability. In Task C, the PA_VCG team achieved the highest performance with an ANLS score of 0.643. Their advanced techniques and fine-tuning processes allowed them to excel in this task, showcasing their model's capability to handle nuanced and challenging OCR scenarios.

We plan to continue this challenge to enrich the literature on Indic handwriting text recognition tasks further. By introducing new methods and datasets, we aim to push the boundaries of what is possible in OCR technology. This ongoing effort will provide valuable benchmarks and stimulate innovation and development within the OCR community.The impact of this challenge extends beyond the immediate competition. It encourages the creation of better models and developing more complex datasets, driving progress in the field. By fostering a competitive yet collaborative environment, we help researchers and developers improve their approaches, leading to more accurate and reliable OCR systems. It, in turn, benefits a wide range of applications, from digital archiving to automated document processing, and contributes to the advancement of technology that can handle diverse and intricate handwriting styles.

Acknowledgement

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

References

1. Easyocr. https://github.com/JaidedAI/EasyOCR/tree/master (2022)
2. Script Grammar. for Indian languages (Accessed March 26 2020), `http://language.worldofcomputing.net/grammar/script-grammar.html.`
3. Baek, Y., Lee, B., Han, D., Yun, S., Lee, H.: Character region awareness for text detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9365–9374 (2019)

4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond (2023)
5. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: European Conference on Computer Vision. pp. 178–196 (2022)
6. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024)
7. Gongidi, S., Jawahar, C.: IIIT-INDIC-HW-WORDs: A dataset for Indic handwritten text recognition. In: ICDAR. pp. 444–459 (2021)
8. Graves, A., Schmidhuber, J.: Offline handwriting recognition with multidimensional recurrent neural networks. In: NIPS (2008)
9. Jemni, S.K., Ammar, S., Kessentini, Y.: Domain and writer adaptation of offline Arabic handwriting recognition using deep neural networks. Neural Computing and Applications (2022)
10. Kenton, J.D.M.W.C., Toutanova, L.K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, p. 2 (2019)
11. Krishnan, P., Jawahar, C.V.: HWNet v2: An efficient word image representation for handwritten documents. IJDAR (2019)
12. Li, M., Lv, T., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F.: TROCR: Transformer-based optical character recognition with pre-trained models. arXiv (2021)
13. Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al.: Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. arXiv preprint arXiv:2311.07575 (2023)
14. Ly, N.T., Nguyen, C.T., Nakagawa, M.: Training an end-to-end model for offline handwritten Japanese text recognition by generated synthetic patterns. In: ICFHR (2018)
15. Maalej, R., Kherallah, M.: Improving the DBLSTM for on-line Arabic handwriting recognition. Multimedia Tools and Applications (2020)
16. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
17. Mindee: doctr: Document text recognition. `https://github.com/mindee/doctr` (2021)
18. Mondal, A., Jawahar, C.: Icdar 2023 competition on indic handwriting text recognition. In: International Conference on Document Analysis and Recognition. pp. 435–453. Springer (2023)
19. Nguyen, K.C., Nguyen, C.T., Nakagawa, M.: A semantic segmentation-based method for handwritten Japanese text recognition. In: ICFHR (2020)
20. Pal, U., Chaudhuri, B.: Indian script character recognition: a survey. Pattern Recognition (2004)
21. Peng, D., Jin, L., Ma, W., Xie, C., Zhang, H., Zhu, S., Li, J.: Recognition of handwritten Chinese text by segmentation: A segment-annotation-free approach. IEEE Transactions on Multimedia (2022)
22. Pham, V., Bluche, T., Kermorvant, C., Louradour, J.: Dropout improves recurrent neural networks for handwriting recognition. In: ICFHR (2014)
23. Tanaka, R., Nishida, K., Yoshida, S.: Visualmrc: Machine reading comprehension on document images. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 13878–13888 (2021)

24. Tito, R., Karatzas, D., Valveny, E.: Document collection visual question answering. In: 16th International Conference on Document Analysis and Recognition (ICDAR). pp. 778–792 (2021)
25. Wu, Y.C., Yin, F., Chen, Z., Liu, C.L.: Handwritten Chinese text recognition using separable multi-dimensional recurrent neural network. In: ICDAR (2017)
26. Xie, Z., Sun, Z., Jin, L., Feng, Z., Zhang, S.: Fully convolutional recurrent network for handwritten Chinese text recognition. In: ICPR (2016)
27. Zhu, F., Lei, W., Feng, F., Wang, C., Zhang, H., Chua, T.S.: Towards complex document understanding by discrete reasoning. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4857–4866 (2022)