ICDAR 2024 Competition on Reading Documents Through Aria Glasses

Soumya Shamarao Jahagirdar[1][0000-0002-3460-9151], Ajoy Mondal[1][0000-0002-4808-8860], Yuheng (Carl) Ren[2], Omkar M Parkhi[2], and C. V. Jawahar[1][0000-0001-6767-7057]

[1] International Institute of Information Technology, Hyderabad, India
[2] Meta Reality Labs Research, Menlo Park, California, United States

Abstract. This paper presents the competition report on Reading Documents through Aria Glasses (ICDAR 2024 RDTAG) held at the 18th International Conference on Document Analysis and Recognition (ICDAR 2024). From a mixed reality perspective, understanding the text in the world is of paramount importance. However, all day long, always on, machine perception devices like Aria Glasses pose a unique primary challenge of lower resolution due to their power and sensor constraints. Moreover, diverse everyday scenes like variations in the lighting conditions and reading positions further complicate the reading tasks. To address this, we propose a new dataset and a challenge. Specifically, we propose three novel tasks: Isolated Word Recognition in Low Resolution (Task A), Prediction of Reading Order (Task B), and Page Level Recognition (Task C). We provide new training and test sets consisting of document images captured by Aria Glasses while reading diverse documents in English under various everyday scenarios. Our aim is to engage researchers with prior experience in English language OCR, and to establish benchmarks contributing to the academic literature in this field. A total of thirty-three different teams from around the world registered for this competition, and twelve teams submitted their results along with algorithm details. The winning team, **SRCB**, achieved a 97.23% Character Recognition Rate (CRR) and a 90.45% Word Recognition Rate (WRR) for Task A: Isolated Word Recognition in Low Resolution. Team **Gang-of-N** won Task B: Prediction of Reading Order with a BLEU score of 0.0939. Team **SRCB** also won Task C: Page Level Recognition and Reading with a 77.44% average Page Level Character Recognition Rate (PCRR) and a 50.55% average Page Level Word Recognition Rate (PWRR).

Keywords: Wearable camera, Aria glasses, word recognition, reading order, page recognition, and reading page.

1 Introduction

The ability to comprehend text within everyday documents is a fundamental task that is necessary to complete numerous tasks humans undertake regularly in their daily lives. The ability to comprehend text is a key pathway to acquire knowledge over time. Similarly, for Contextual AI systems, there is a need to develop solutions to give machines the ability to read and infer from life-long activities and develop skills. Life-long comprehension requires long term always on data capture, which introduces a number of challenges to the sensor configuration (cost, sensor, energy efficiency, how to wear, etc.) and the usage pattern (human pose, reading gestures, etc.). Combined with the task of text comprehension, these bring a unique challenge for the ICDAR community. We introduce this challenge with the help of Project Aria[3] from the Meta Reality Labs Research.

In addition to the design challenges discussed above, in wearable device document comprehension, further complexities arise from the variability of the human poses – ranging from inclined or resting positions to even moments of sleep – coupled with varying lighting conditions, encompassing sunlight, artificial lamps, or night-time settings. Additionally, potential obstacles such as occlusions due to page folding or other subject in the scene pose additional hurdles in obtaining accurate OCR. The diverse nature of document types further increases the intricacy and variability of the task at hand. Various categories ranging from textbooks to academic dissertations, newspapers to conference papers, and encyclopedias to biographies, each presenting a unique challenge in text recognition. The distinctive layouts, font styles, and content structures inherent in newspaper editorials, research periodicals, dictionaries, and others demand adaptable OCR systems that are capable of handling these different formats.

We believe that OCR will be a key technology piece that needs to be solved for EgoCentric Machine Perception. It brings its unique challenges, as discussed above. While we think that some of the sensor constraints would get relaxed over the coming years, challenges of egocentric viewpoints are here to stay. We believe that with Project Aria, we are at a point to begin the journey of EgoCentric OCR. With this context in mind, in this competition, we aim to introduce the task of low-resolution OCR on pages captured using wearable devices, focusing on the complex challenges posed by diverse document types and the complexities of varying human positions and lighting conditions. The following are the tasks that the competition would look into:

1. Task A: Isolated Word Recognition in Low Resolution
2. Task B: Prediction of Reading Order
3. Task C: Page level Recognition and Reading

2 Dataset

We have curated a diverse dataset of raw document images captured by Aria Glasses, encompassing various English content genres. The dataset comprises documents such as Textbooks, Newspapers, and Magazines. The document images are captured while reading under a spectrum of indoor and outdoor situations, covering different times of day and night. The diverse lighting conditions

[3] https://www.projectaria.com/

include daytime and nighttime. This comprehensive approach ensures the inclusion of a wide range of scenarios that may be encountered in real-world reading situations. The dataset encompasses near to 2K images. All images in the dataset undergo meticulous manual annotation to provide detailed information for analysis. The annotations include (i) bounding boxes: precise annotation of word-level bounding boxes on each page, (ii) reading order: Annotation of the reading order, capturing the natural flow of content on the page, and (iii) text transcription: comprehensive text transcription of words present on each page, providing ground truth for content recognition. The annotations are structured in XML format, facilitating easy integration into various recognition and analysis pipelines. We create an RDAG-1.0 dataset consisting of near to 2K page images. The dataset is divided into a training set of 1600 page images and test set of 363 images. We only provide ground truths for the training set. Fig. 1 shows a few sample images captured by Aria glasses under various conditions.

3 Evaluation Metric

Task A: Isolated Word Recognition in Low Resolution: 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

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

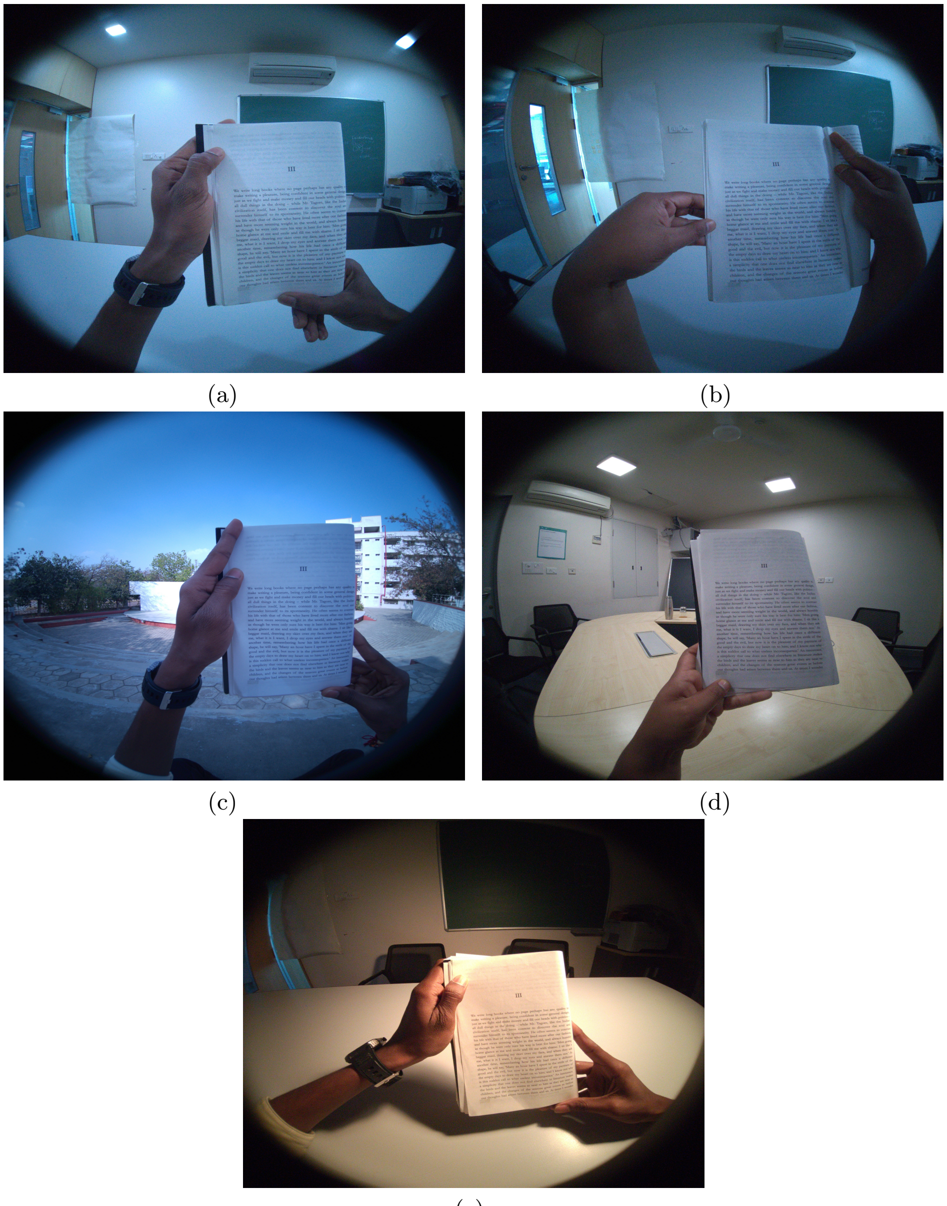
where S indicates the number of substitutions, D indicates the number of deletions, I indicates the number of insertions, and N number of instances in reference text. In the case of CER, Eq. 1 operates on character level, and in the case of WER, Eq. 1 operates on word level. Recognition Rate (RR) is 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: Prediction of Reading Order: We use an average page level BLEU score for the evaluation of the reading order prediction task. BLEU [10] is widely used in sequence generation tasks. Since reading order prediction is a sequence-to-sequence mapping, it is natural to evaluate the performance of an algorithm for reading order prediction with BLEU scores. BLEU scores measure the n-gram overlaps between the hypothesis and reference. We define the average page level BLEU score for this task. The page level BLEU refers to the micro-average precision of n-gram overlaps within a paragraph.

Task C: Page Level Recognition: 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).



(a)

(b)

(c)

(d)

(e)

Fig. 1. The images display various samples captured by Aria glasses, each depicting a book page under five different conditions: (a) indoors close to a light source, (b) indoors far from a light source, (c) outdoors, (d) indoors at night under white light, and (e) indoors at night under yellow light.

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

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.

4 Methods

Thirty-three participants around the world registered for the competition. However, we got submissions from twelve of them. These twelve teams are - (i) W2024, Capital Normal University, China (ii) ONEPINGAN, PING AN LIFE INSURANCE (GROUP) COMPANY OF CHINALTD, China, (iii) DXM-DI-AI-CV-TEAM, Duxiaoman, (iv) Gang-of-N, PING AN LIFE INSURANCE (GROUP) COMPANY OF CHINALTD, China, (v) TeamOCR, IIIT Nagpur, India, (vi) IndicOCR Group, Digital University Kerala, India, (vii) SRCB, Ricoh Software Research Center Beijing Co., Ltd, China, (viii) ljw2333, Beijing Institute of Technology, China, (ix) Go Crazy, tsinghua university, China, (x) ZYFGali, Private, (xi) TSNUK, Taras Shevchenko National University of Kyiv, Ukraine, (xii) Independent OCR, NIT Surat, India.

4.1 Methods for Task A

DXM-DI-AI-CV-TEAM: CLIP4STR [13], a simple yet effective STR method built upon image and text encoders of CLIP. CLIP can perceive and understand text in images, even for irregular text with noise, rotation, and occlusion. CLIP is potentially a powerful scene text recognition expert. It has two encoder-decoder branches: a visual branch and a cross-modal branch. The visual branch provides an initial prediction based on the visual feature, and the cross-modal branch refines this prediction by addressing the discrepancy between the visual feature and text semantics. To fully leverage the capabilities of both branches, the team designed a dual predict-and-refine decoding scheme for inference. The pre-trained model is then then fine-tuned with the required training set.

Go Crazy: In this experiment, the team adopt the CLIP-OCR [12] algorithm based on pure vision. This method explore the potential of the Contrastive Language-Image Pre-training (CLIP) model in scene text recognition (STR), and establish a novel Symmetrical Linguistic Feature Distillation framework (named CLIP-OCR) to leverage both visual and linguistic knowledge in CLIP. Different from previous CLIP-based methods mainly considering feature generalization on visual encoding, CLIP-OCR proposes a symmetrical distillation strategy (SDS) that further captures the linguistic knowledge in the CLIP text encoder. By cascading the CLIP image encoder with the reversed CLIP text encoder, a symmetrical structure is built with an image-to-text feature flow that covers not only visual but also linguistic information for distillation. The team first trained a basic model based on publicly available scene text datasets and then fine-tuned the model on the required training set to get the final model.

ljw2333: The team use CLIP4STR [13] for this experiment. The pre-trined model with scene text datasets is then fine-tuned with required training set to get final model.

SRCB: The method PARSeq [3], learns an ensemble of internal AR LMs with shared weights using Permutation Language Modeling. It unifies context-free non-AR and context-aware AR inference, and iterative refinement using bidirectional context. And PARSeq is optimal on accuracy vs parameter count, FLOPS, and latency because of its simple, unified structure and parallel token processing.

Data Generation: (i) Collect representative background images from the test set images; (ii) Use text_renderer to synthesize data (use different probabilities to combine different ways of augmenting data).

Training Data Pre-processing: (i) Train a model using synthetic and exposed data; (ii) Use the trained model to predict the training set data; (iii) Each prediction result is compared with the corresponding label, and if the prediction result is inconsistent with the corresponding label and the score of the prediction result is greater than 0.95, the label is replaced with the prediction result.

Additional Training Data: (1) Synthetic data: 1.2M and (2) SROIE(ICDAR2019): 33K.

Training Strategy: (1) Use pre-trained model to train the model with all training dataset; (2) Use Training Data Pre-processing steps 2 and 3 to correct the training dataset; Repeat steps 1 and 2 about 3 times and save the best model.

Post OCR Error Correction: It is found that due to a certain error rate of word detection coordinates, sometimes the letters of the previous word or the next word will appear in the front or back position of the word picture, and the model will recognize these letters, which will lead to some more characters that should not exist in the final result of recognition, so the team did a simple post-processing. 1) If the recognition result contains a letter and a punctuation mark, the recognition result retains only the punctuation marks; 2) If the recognition result contains spaces, only the content after the spaces is retained in the recognition results.

W2024: CLIP4STR [13] is a straightforward yet highly effective Scene Text Recognition (STR) method that leverages CLIP's image and text encoders. CLIP can perceive and comprehend text within images, even when irregular, noisy, rotated, or occluded, making it a potentially powerful tool for scene text recognition. CLIP4STR employs two encoder-decoder branches: a visual branch and a cross-modal branch. The visual branch provides an initial prediction based on visual features, while the cross-modal branch refines this prediction by addressing the differences between visual features and text semantics. To maximize the potential of both branches, we have designed a dual predict-and-refine decoding scheme for inference. The team pre-trained the model with available scene text recognition datasets and then fine-tuned with required training set.

ZYFGali: In this experiment, the team adopt the SVTR [5] recognition algorithm based on pure vision. Unlike the traditional extraction of features, the SVTR algorithm introduces local and global hybrid blocks to extract stroke features and inter-character correlation respectively, and combines multi-scale backbone to form multi-granularity feature description. The team first trained a basic model based on publicly available scene text datasets and then fine-tuned the model on the required training set to get the final model.

IndependentOCR: This team used PARSeq [3] for this experiments. They used printed English word level images to pre-trained the model. After that they fine-tuned the model with required training set.

TeamOCR: This team employs the CRNN architecture [6], which comprises four main modules: the Transformation Network (TN), Feature Extractor (FE), Sequence Modeling (SM), and Predictive Modeling (PM). The Transformation Network consists of 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 utilizes the ResNet architecture, while the Sequence Modeling module features a 2-layer Bidirectional LSTM (BLSTM) with 256 hidden neurons per layer. Finally, the Predictive Modeling module employs Connectionist Temporal Classification (CTC) for character decoding and recognition, aligning the feature sequence with the target character sequence.

4.2 Methods for Task B

Go Crazy: The team trained the layoutreader [11] model using word text, word box, and word order in the training set to construct the data. The team used the layoutreader pre-trained model[4].

Gang-of-N: This team integrates the least squares method with the LayoutReader [11] framework to predict the correct reading order. LayoutReader introduces a novel system for reading and comprehending document layouts by

[4] `https://huggingface.co/nielsr/layoutreader-readingbank`

combining natural language processing (NLP) and computer vision techniques. The least squares method is employed to restore the slope of Chinese characters across the entire image, facilitating an overall prediction of the reading order. Meanwhile, the LayoutReader framework addresses local reading order issues, such as those found in multi-column texts.

Preprocessing: Construction of the training set: The team randomly shuffles the text in the images from the training set five times to create augmented training data, which is then converted into JSON format. Construction of the test set: The test dataset is converted into JSON format compatible with the LayoutReader model, which may involve (i) data format transfer and (ii) random shuffling.

Postprocessing The team combines the results of all test sets from the 10-fold cross-validation and the least squares method.

ONEPINGAN: This team combines the least squares method and the LayoutReader [11] scheme to predict the correct reading order. LayoutReader is a research paper that presents a new system designed for reading and understanding the layout of documents. This system combines both natural language processing (NLP) and computer vision techniques to predict the reading order. The purpose of the least squares method is to restore the slope of the Chinese characters in the entire image to achieve an overall prediction of the reading order. The layout reader scheme addresses the local reading order issues, such as multi-column texts.

Preprocessing: Construction of the training set: the team randomly shuffles the text in the images from the training set five times to create our augmented training data and convert them into JSON format. Construction of the test set: We convert the test dataset into JSON format compatible with the layout reader model. This might include: (i) Data formate transfer, and (ii) Randomly shuffle.

Postprocessing The team combines the results of all test sets from the 10-fold cross-validation and the least squares method.

SRCB: The main idea of solution is recognizing the text lines and classify the bounding box of layout unit like paragraph. As test set of Task B does not provide the image of the original page, existed text detection methods like DBNet is not suitable in this task. So, the team builds a black box image with the bounding box of each word in the page and train a semantic segmentation model (PaddleSeg[5]) to recognize the link of the middle point of the word boxes in the same line. The training data of this segmentation model was constructed from the training data. The team uses the bounding boxes in the training set to build the black box image and use the relative position between order neighboring boxes

[5] `https://github.com/PaddlePaddle/PaddleSeg/tree/release/2.9`

to judge whether two boxes are in the same line. With the segmentation result, we use morphological processing of images, such as dilation and erosion, on it to get the layout unit like paragraph. And the reading order of the paragraphs and text lines in one paragraph are mainly decided with rules like from the top to the bottom and from the left to the right.

W2024: The team has trained three models in total, which are the layout analysis model, the word-level reading order prediction model, and the block level reading order prediction model. The train flow of the proposed method is as follows: (i) Since the test set only contains word text and word box, there is no corresponding image. Therefore, the team uses all the word text and their box from the labels in the training set to render the images on a blank sheet of paper, obtaining a total of 1600 rendered images, and then labeling their text blocks with rectangles. This data is then used to train a layout analysis model based on pre-trained layoutlmv3 [8][6]. (ii) Use the layout analysis model to infer 1600 rendered images from training set and obtain a total of 6423 blocks. Then use the IoU overlapping relationship between the boxes of these blocks and the word boxes to classify the words into corresponding blocks, thus obtaining the word content and word boxes in 6423 blocks. (iii) The layoutreader model was trained using the word content, word boxes and their sorted labels in the 6423 blocks to obtain a model that could get the order in the blocks. (iv) Train the other layoutreader model in blocks to obtain the model that can get the order between blocks, and use some rules to improve the ability of ordering between blocks. For example, the center coordinates of all block boxes are used to determine whether the text in the image is single or double columns. (v) The team used the pre-trained layoutreader model[7].

4.3 Methods for Task C

SRCB: The main idea for solution of Task C is based on the text recognition in Task A and reading order in Task B. The main process of the method is (i) detect the page area from to photo, (ii) recognize the words with their bounding boxes in the pages, and (iii) format the reading order.

For page detection, the team trained an object detection model with the training data. The label of the detection objection was constructed from the minimum and maximum of the x and y coordinates of all the boxes in the image. For word detection, the team trained a detection model DBNet with the bounding boxes of the words in the training data. Word recognition and reading order parts are same in the Task A and Task B.

W2024: The team has trained two models in total, namely the book detection model and the whole page recognition model. The train flow of our method is as follows: (i) the team used rectangular boxes to label the book borders of 1600

[6] `https://huggingface.co/microsoft/layoutlmv3-base`

[7] `https://huggingface.co/nielsr/layoutreader-readingbank`

images in the training set, which were used to train the YOLOV4 [4]-based object detection model. The pre-trained model, the team used comes from[8] (ii) The team use the trained object detection model to infer 1600 training set images, get the bounding box of the book, and then crop the book out. The cropped images and labels are used to train the nougat model, whose pre-trained model is from[9].

The inference flow is as follows: (i) The object detection model is used to infer the original image, and then the detection results are used to crop the original image. (ii) The nougat model is used to infer the cropped image, and the recognition result is obtained.

IndependentOCR: The method follows three steps - (i) detect book regions in image using Mask R-CNN model [7], (ii) then detect individual words in the page using CRAFT [2], and (iii) finally each detected word is recognized by PARSeq [3].

TeamOCR: The method follows two steps - (i) detect individual words in the page using DocTR [9], and (ii) then each detected word is recognized by CRNN [6].

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

5 Competition Results

5.1 Results Analysis for Task A

Quantitative Results: — Among the thirty-three registered participants, ten teams submitted results for Task A. Table 1 displays the obtained Character Recognition Rate (CRR) and Word Recognition Rate (WRR) for Task A across different teams. The **SRCB** team achieved the highest scores, with a CRR of 97.23% and a WRR of 90.45%. They utilized PARSeq, incorporating additional data for pre-training the model, fine-tuning with the required training set, and applying post-processing to enhance accuracy. The runner-up, team Go Crazy, employed a pre-trained CLIP-OCR model with scene text recognition datasets and then fine-tuned it with the necessary training set. They achieved notable results, though slightly behind the SRCB team. Teams W2024 and ljw2333 used CLIP4STR for this task, achieving CRRs of 92.32% and 93.57% and WRRs of 81.56% and 83.26%, respectively. Both teams showed competitive performance but could not surpass the top two. The IndicOCR Group had the lowest performance, indicating potential areas for improvement in their approach. Overall, the results highlight the effectiveness of pre-training and fine-tuning strategies in achieving high recognition rates.

[8] `https://github.com/bubbliiiing/yolov4-pytorch/releases/download/v1.0/yolo4_weights.pth`

[9] `https://github.com/facebookresearch/nougat`

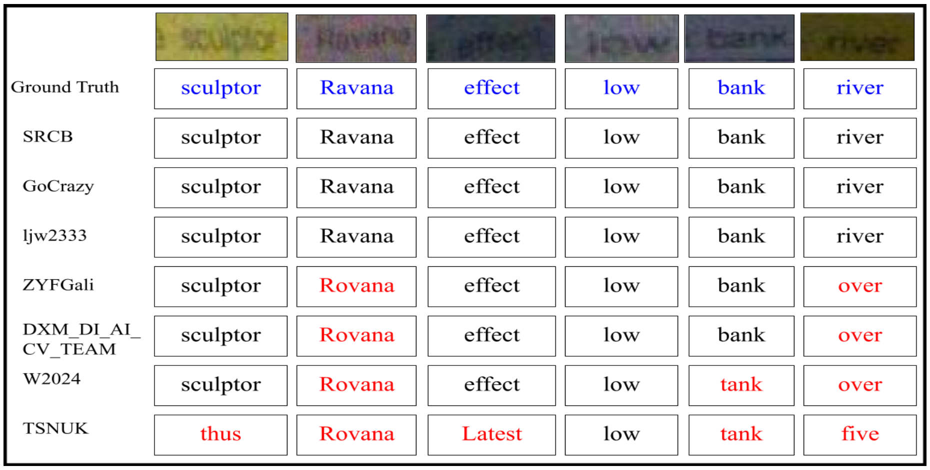
Team	**CRR**	**WRR**
SRCB	**97.23**	**90.45**
Go Crazy	95.25	84.65
ljw2333	93.57	83.26
ZYFGali	95.08	83.23
DXM-DI-AI-CV-TEAM	92.32	81.56
W2024	92.32	81.56
TSNUK	74.18	63.23
IndependentOCR	72.58	62.73
TeamOCR	68.23	43.21
IndicOCR Group	65.78	40.53

Table 1. Shows CRR and WRR for Task A of different teams. The bold value indicates the best results.

Qualitative Results: — Fig. 2 showcases several samples of qualitative results obtained by various teams. The SRCB, GoCrazy, and ljw2333 teams achieved correct predictions for all sample word-level images, demonstrating their models' robustness in handling low-resolution word level images. It indicates their models' ability to interpret and recognize text effectively despite the challenging image quality. In contrast, the ZYFGali and DXM_DI_AI_CV_TEAM teams made incorrect predictions for two samples each, due to characters not being visible properly and ambiguities caused by low resolution. The team w2024 predicted three incorrect words, while the team TSNUK predicted five incorrect outputs out of six words, highlighting difficulties in their models' accuracy. The SRCB team achieved high performance by using additional synthetic and real datasets for training and employing post-processing techniques to correct OCR outputs. While GoCrazy, ljw2333, and DXM_DI_AI_CV_TEAM all used CLIP4STR, their performances were very close to each other, showcasing the competitive nature of their approaches. This comparison underscores the importance of comprehensive training data and robust post-processing in enhancing OCR model performance.

5.2 Results Analysis for Task B

Quantitative Results: — Seven teams submitted results for Task B and Table 2 shows quantitative results of all teams. Among them, the Gang-of-N team achieved the best performance with a BLEU score of 0.0939, showcasing their superior approach. The W2024 team followed as the runner-up with a BLEU score of 0.0829, indicating a competitive but slightly less effective method. The IndicOCR Group had the lowest performance, scoring a BLEU of 0.0271, highlighting significant challenges in their approach. Due to the shuffling of words in the test set, all methods exhibited relatively low performance. This complexity made accurate sequence prediction particularly challenging. However, Gang-of-N stood out by integrating the least squares method with the LayoutReader approach, which allowed them to achieve the best results despite the difficult conditions. The W2024 team's close second-place finish suggests their method was robust



Ground Truth	sculptor	Ravana	effect	low	bank	river
SRCB	sculptor	Ravana	effect	low	bank	river
GoCrazy	sculptor	Ravana	effect	low	bank	river
ljw2333	sculptor	Ravana	effect	low	bank	river
ZYFGali	sculptor	Rovana	effect	low	bank	over
DXM_DI_AI_ CV_TEAM	sculptor	Rovana	effect	low	bank	over
W2024	sculptor	Rovana	effect	low	tank	over
TSNUK	thus	Rovana	Latest	low	tank	five

Fig. 2. shows visual results obtained by various teams. Blue colored text indicates ground truths, Red colored text indicates wrongly recognized text.

but couldn't match Gang-of-N's effectiveness. The substantial gap between the top performers and the IndicOCR Group underscores the difficulty of the task and the importance of advanced methodologies. The generally low BLEU scores across all teams emphasize the inherent challenges of Task B, mainly due to the shuffled word sequences. These results suggest further innovation and refinement in OCR technologies to handle such complex scenarios better.

Team	**BLEU**
Gang-of-N	**0.0939**
W2024	0.0829
GoCrazy	0.0792
SRCB	0.0741
IndependentOCR	0.0541
TeamOCR	0.0431
IndicOCR Group	0.0271

Table 2. Shows BLEU score for Task B of different teams. The bold value indicates the best results.

5.3 Results Analysis for Task C

Quantitative Results: — Table 3 presents the PCRR (Page Character Recognition Rate) and PWRR (Page Word Recognition Rate) scores for different teams participating in Task C. Among the five teams, SRCB achieved the highest scores, with a PCRR of 77.44% and a PWRR of 50.55%. The SRCB team employed a three-step process for this task, contributing significantly to their superior performance. Despite their success, various capturing conditions posed challenges, making it difficult for methods to recognize text correctly. These conditions included lighting, angle, and image quality variations, which impacted the accuracy of text recognition across all teams. On the other end of the spectrum, the IndicOCR Group was the least-performing team, with a PCRR of 30.78% and a PWRR of 12.58%. Their results highlight their difficulties adapting their methods to the diverse capturing conditions. The score disparity between SRCB and IndicOCR Group emphasizes the importance of robust preprocessing and post-processing techniques in improving recognition rates. Additionally, these results suggest that teams need to develop more adaptable and resilient models to handle the complexities of real-world image capture scenarios effectively.

Team	**PCRR**	**PWRR**
SRCB	77.44	50.55
W2024	42.29	26.77
IndependentOCR	39.46	24.84
TeamOCR	36.92	21.75
IndicOCR Group	30.78	12.58

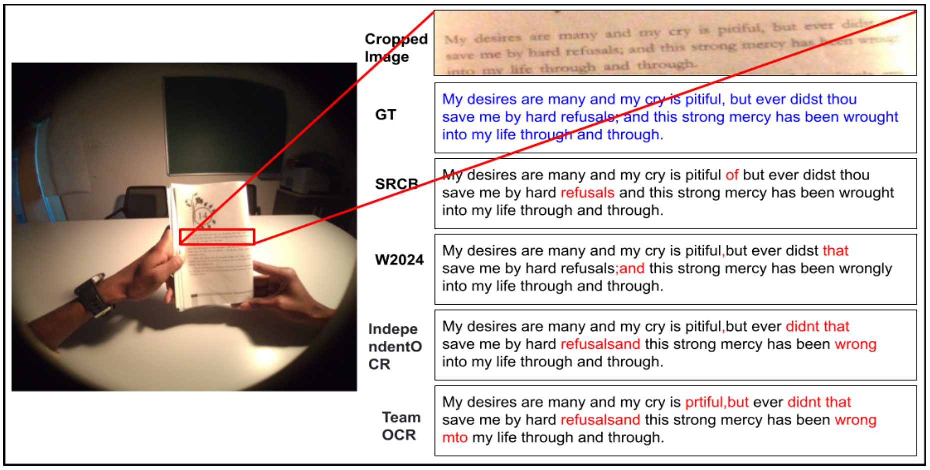
Table 3. Shows average Page Level CRR (PCRR) and average Page Level WRR (PWRR) for Task C of different teams. The bold value indicates the best method.

Qualitative Results: — Figs. 3 and 4 display sample outputs from several teams. These figures illustrate the varying levels of accuracy achieved by different models in recognizing text under diverse conditions. The outputs highlight the strengths of the top-performing teams, such as SRCB, whose results consistently show high precision in text recognition.

For instance, SRCB's samples demonstrate their model's ability to accurately interpret and process text even in challenging scenarios, reflecting their effective three-step process. In contrast, samples from lower-performing teams, like IndicOCR Group, reveal difficulties in handling complex backgrounds and varying lighting conditions, leading to lower accuracy rates. These visual comparisons underscore the importance of advanced pre-processing and post-processing techniques. They also emphasize the need for robust training datasets encompassing many real-world conditions to improve model performance. The sample outputs serve as a clear indicator of each team's approach and the effectiveness of their methodologies in addressing the challenges posed by Task C.

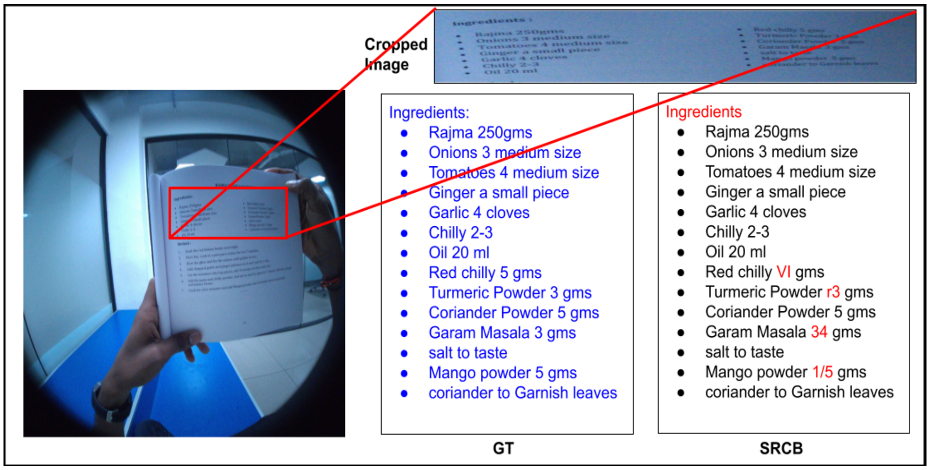
6 Conclusion

This competition motivates researchers with prior experience in English language OCR to establish benchmarks and contribute to academic literature in


Cropped Image
My desires are many and my cry is pitiful, but ever didst thou save me by hard refusals; and this strong mercy has been wrought into my life through and through.

GT	My desires are many and my cry is pitiful, but ever didst thou save me by hard refusals; and this strong mercy has been wrought into my life through and through.
SRCB	My desires are many and my cry is pitiful of but ever didst thou save me by hard refusals and this strong mercy has been wrought into my life through and through.
W2024	My desires are many and my cry is pitiful,but ever didst that save me by hard refusals;and this strong mercy has been wrongly into my life through and through.
IndependentOCR	My desires are many and my cry is pitiful,but ever didnt that save me by hard refusalsand this strong mercy has been wrong into my life through and through.
Team OCR	My desires are many and my cry is prtiful,but ever didnt that save me by hard refusalsand this strong mercy has been wrong mto my life through and through.

Fig. 3. Show qualitative results of several teams on a page image captured by Aria glass.



GT	SRCB
Ingredients:	Ingredients
Rajma 250gms	Rajma 250gms
Onions 3 medium size	Onions 3 medium size
Tomatoes 4 medium size	Tomatoes 4 medium size
Ginger a small piece	Ginger a small piece
Garlic 4 cloves	Garlic 4 cloves
Chilly 2-3	Chilly 2-3
Oil 20 ml	Oil 20 ml
Red chilly 5 gms	Red chilly VI gms
Turmeric Powder 3 gms	Turmeric Powder r3 gms
Coriander Powder 5 gms	Coriander Powder 5 gms
Garam Masala 3 gms	Garam Masala 34 gms
salt to taste	salt to taste
Mango powder 5 gms	Mango powder 1/5 gms
coriander to Garnish leaves	coriander to Garnish leaves

Fig. 4. Show qualitative results of the best performing team SRCB on a page image captured by Aria glass.

this field. Only twelve of the thirty-three registered teams submitted their results and detailed their algorithms. In Task A, the SRCB team emerged victorious with a Character Recognition Rate (CRR) of 97.23% and a Word Recognition Rate (WRR) of 90.45%. The GoCrazy team was the runner-up, achieving a CRR of 95.25% and a WRR of 84.65%. SRCB's success was attributed to its use of additional synthetic and real training data and effective post-processing techniques. For Task B, the Gang-of-N team secured the top position with a BLEU score of 0.0939, while the W2024 team followed closely as the runner-up. Gang-of-N achieved the best results by integrating the least squares method with LayoutReader to predict sequences, demonstrating their innovative approach.In Task C, SRCB again led the competition, achieving a Page Level Character Recognition Rate (PCRR) of 77.44% and a Page Level Word Recognition Rate (PWRR) of 50.55%. The W2024 team came in second, with a PCRR of 42.29% and a PWRR of 26.70%. The competition highlights the significant advancements and ongoing challenges in OCR technology, particularly under low-resolution conditions. We plan to continue this challenge to enhance the literature on text recognition with new methods and datasets. This initiative significantly impacts the OCR community by fostering the development of better models and more complex datasets, pushing the boundaries of what OCR systems can achieve in real-world applications.

Acknowledgement

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

References

1. Tesseract ocr. https://github.com/tesseract-ocr/tesseract (2021)
2. 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)
3. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: European conference on computer vision. pp. 178–196. Springer (2022)
4. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
5. Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., Du, Y., Jiang, Y.G.: Svtr: Scene text recognition with a single visual model. arXiv preprint arXiv:2205.00159 (2022)
6. Gongidi, S., Jawahar, C.: IIIT-INDIC-HW-Words: A dataset for indic handwritten text recognition. In: ICDAR. pp. 444–459 (2021)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
8. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 4083–4091 (2022)

9. Mindee: doctr: Document text recognition. `https://github.com/mindee/doctr` (2021)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
11. Wang, Z., Xu, Y., Cui, L., Shang, J., Wei, F.: Layoutreader: Pre-training of text and layout for reading order detection. arXiv preprint arXiv:2108.11591 (2021)
12. Wang, Z., Xie, H., Wang, Y., Xu, J., Zhang, B., Zhang, Y.: Symmetrical linguistic feature distillation with clip for scene text recognition. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 509–518 (2023)
13. Zhao, S., Quan, R., Zhu, L., Yang, Y.: Clip4str: A simple baseline for scene text recognition with pre-trained vision-language model. arXiv preprint arXiv:2305.14014 (2023)