ICPR 2024 Competition on Word Image Recognition from Indic Scene Images

Harsh Lunia[0009−0007−4155−2011], Ajoy Mondal[0000−0002−4808−8860], and C. V. Jawahar[0000−0001−6767−7057]

Centre for Vision Information Technology,
International Institute of Information Technology, Hyderabad 500032, India
harsh.lunia@research.iiit.ac.in
{ajoy.mondal,jawahar}@iiit.ac.in

Abstract. Scene text recognition has historically concentrated on English, with limited advancements in developing solutions that perform well across multiple languages. Previous efforts in multilingual scene text recognition have predominantly targeted languages with considerable syntactic and semantic differences. However, Indian languages, while diverse, share numerous common features that remain largely underutilized. This competition aims to address the often-overlooked challenge of scene text recognition within the Indian context and to advance robust word image recognition across ten Indian languages.
The dataset provided for this competition is one of the most comprehensive multilingual datasets, encompassing 10 languages, each with 17,500 training samples, 2,500 validation samples and 5,000 test word-image samples. The task was to correctly recognize the word-images, for which we received forty-nine registrations and five final submissions from industrial and research communities. The winning team achieved an average Character Recognition Rate (CRR) of 92.85% and a Word Recognition Rate (WRR) of 84.01% across the ten languages. This paper details the proposed dataset and summarizes the submissions for the competition—WIRIndic-2024.

Keywords: OCR · Scene Text Recognition · Indic/Indian Languages.

1 Introduction and Related Work

Rich textual content in natural settings contains valuable information that significantly enhances understanding of the environment in contemporary times. Such text is utilized for tasks like image search, translation, transliteration, assistive technologies (especially for the visually impaired), autonomous navigation, and more.

Scene text recognition is increasingly crucial, with its solution promising advancements in various downstream tasks like the above. Despite significant progress in this field, there are areas for further enhancement. These include developing models that can handle diverse languages, fonts, layouts, and styles

and creating solutions resilient to text-related image imperfections such as blurriness, occlusion, and uneven illumination. Researchers have sought to tackle these challenges by curating datasets tailored to specific problems, each highlighting distinct features and representing subsets of real-world challenges.

Competitions have also been organized, each addressing specific subsets of the above-mentioned challenges. For instance, [4] focused on addressing arbitrary-shaped scene text instances. Additionally, [5] compiled a comprehensive collection of public scene text datasets across seven datasets to enhance solutions' robustness to out-of-vocabulary words. The collection ensured that both test and validation splits included at least one scene text instance outside the training set's vocabulary. Recognizing the need for solutions tailored to road text, which is often dispersed and subject to distortions like motion blur, [19] organized the *RoadText* Competition, focusing on video text detection and recognition. Following the demonstration by [11] that the unification of OCR and geometric layout analysis, termed as *Hierarchical Text Detection and Recognition (HTDR)*, benefits both tasks, [12] hosted a competition on *Hierarchical Text Detection and Recognition*. In this competition, participants were expected to perform text detection, recognition, and layout analysis. Likewise, to enable scene text solutions across multiple languages, the *MLT-17 robust reading challenge* [17] and its subsequent iteration in 2019 [16] were tailored explicitly for this purpose. These challenges marked the community's initial endeavour to address the multi-lingual setting comprehensively. They offered a curated dataset comprising 20,000 scene images in ten languages, supplemented by synthetically generated images of the same languages to aid training. However, it's worth noting that the targeted languages encompass diverse linguistic and geographical groups.

Fig. 1: Hindi word-images with ground truth (top) and model predictions (bottom, errors in red). Issues highlighted include data scarcity affecting visual learning, high n-gram variability complicating linguistic understanding, challenges in balancing visual and linguistic information, and difficulties with long or complex texts.

Because Indian language scripts are visually more complex and have much larger output space than English, not all Latin STR models can mimic performance in Indian STR solutions [15]. Although spoken by 17% of the world's population, STR solutions for Indian languages remain underdeveloped. Key challenges include a lack of real datasets, which limits the model's ability to learn accurate visual representations of characters and significant variability in word-level n-grams, complicating linguistic feature learning. In Fig 1 the first row shows that while the Latin SOTA model [2] trained on Hindi scene text data can recognize in-vocabulary words, it struggles with out-of-vocabulary sample (4^{th} sample) and exhibits linguistic pre-training biases that override visual cues, leading to misinterpretation of character sequences (3^{rd} sample). The second row highlights further difficulties with long texts and multi-word sequences, causing recognition errors. These issues emphasize the need for enhanced visual feature learning and better integration of linguistic and visual information. Non-Latin languages have made less progress, and existing Latin STR models need to generalize better to different languages [3]. We aim to tackle this challenge through this competition within the scene text domain of the Document Analysis and Recognition community.

We released a comprehensive dataset for Indic languages, comparable in scale to Latin script datasets, encompassing 250K word-image samples across 10 languages. Such a large-scale dataset for Indian languages had been lacking, hindering intensive and serious solution development in the Indic space. The languages included in the dataset are Bengali, Kannada, Hindi, Telugu, Gujarati, Malayalam, Marathi, Odia, Punjabi, and Tamil. Participants were encouraged to utilize additional datasets, both synthetic and real, to develop more robust and high-performing scene text recognition solutions.

2 Organization

This competition comprised only one task - recognition of text in word images (see Section 4.1) extracted from Indic scene images. As part of the competition, we focused solely on recognition, aiming to aggressively develop a robust and more generalized STR solution. The dataset developed for this challenge, detailed in Section 3, provides a solid foundation for training models. Participants were also permitted to enhance their model training by using additional datasets, including synthetic ones.

The competition was hosted on a web portal `https://ilocr.iiit.ac.in/icpr_2024_wirindic/`, which facilitated participant interaction, provided challenge information, registration links, schedules, download links, online submissions, and real-time leaderboard viewing. **Forty-nine** registrations were received, with **five teams** making one or more submissions after the test set was released 13 days before the submission deadline. Some participants submitted results for all ten languages, while others submitted for a subset. In cases of multiple submissions, the highest-performing result for each language was considered, and the overall score was calculated accordingly.

The top three solutions for each language were awarded 5, 4, and 3 points based on WRR and CRR rankings (as described in Section 4.2). All other submissions received 1 point each. Points were allocated separately for CRR and WRR and summed independently for each criterion. The final score for each team was the total points across all submitted languages. Awards were given in two categories: one for WRR and another for CRR, with the teams earning the highest total points in each category declared the winners.

3 Indic Scene Text Dataset

Table 1: Comparison of real scene text dataset sizes: Our proposed dataset, highlighted here, is not only the largest among Indic STR datasets but also rivals or surpasses the sample count of Latin real-world datasets.

	Dataset	# Word Images	Languages	Features
Mono-Lingual	IIIT5K	5K	EN	Reg
	SVT	647	EN	Reg
	IC15	2077	EN	Irreg, Blur
	CUTE80	288	EN	Irreg, LR
	SVTP	1095	EN	Irreg
	COCO-Text	83K	EN	Irreg
	Uber-Text	285K	EN	Irreg
	Urdu-Text	14K	UR	Irreg
Multi-Lingual	LSVT	400K	ZH, EN	Irreg
	MLT-19	89K	AR, BN, HI, ZH ... ($n = 10$)	Reg
	MTWI	289K	ZH, EN	Irreg
	IndicSTR12	27K	ML, HI, BN, UR ... ($n = 12$)	Irreg, LR, Occ
	Ours	250K	ML, HI, BN, TA ... ($n = 10$)	Irreg, LR, Occ

Addressing the scarcity of annotated real data for Indian languages has been a significant challenge. The difficulties in collecting real datasets for Indic scene text are twofold: (i) In India, the coexistence of English and Indian languages in natural settings makes it challenging to gather large-scale datasets exclusively for Indian languages; (ii) Although these languages are widely spoken due to India's large population, they are often geographically concentrated. While popular, India's diverse scripts and languages are used in relatively confined areas, making large-scale data collection more complex than English or other Latin-based languages, which share more significant script commonality and usage over a much larger region. However, recent efforts have led to the development of usable datasets (see Table 1). The MLT-17 [17] and MLT-19 [16] datasets were pioneers, targeting multiple languages, including Bengali and Hindi. More recently, Lunia *et al.* [13] introduced a comprehensive dataset specifically for the Indic space, encompassing 12 languages sourced from Google images. Though the datasets are publicly available, they are limited for training purposes.

Fig. 2: Top Image: shows a few sample captured images. Bottom Image: word-level images extracted from source images.

We curate the largest dataset to solve scene text data scarcity in Indic languages. We sourced our dataset from Indian road scenes, leveraging their text-rich elements, including shop boards, advertisement hoardings, traffic signage boards, banners, pamphlets, and house plates.

3.1 Scene Text Image Collection and Annotation

Fig. 3: **Dataset Annotation Process**: Three stages of the annotation process for the proposed Indic scene text datasets. The first image depicts data collection using GoPro cameras mounted on cars to capture Indian scene images across different cities. The second image shows the use of an automated pipeline for initial word sample detection and text prediction. The third image highlights human annotators verifying and refining the generated annotations and labels.

As shown in Fig 3, the dataset collection and annotation process followed three stages. First, cars equipped with GoPro cameras were driven through various Indian states to capture roadside images. Second, an automated pipeline

processed these images, identifying frames with significant text samples. Pre-trained models performed initial annotations by detecting word instances and predicting corresponding text. Finally, human annotators meticulously verified the annotations using a four-corner point methodology, ensuring accurate labelling of both horizontal and curved word structures, as in [17]. Annotators followed reading directions and minimized background inclusion while validating predicted text labels. We release word-images from Indian road scenes with verified labels, as shown in Fig. 2.

3.2 Dataset Analysis

Fig. 4: Varieties of irregular word image samples: Clockwise from the top-left, the figure showcases occluded word level images, low-resolution samples, font variations, multi-oriented samples, illumination variations, and perspective text. These diverse samples encompass various Indian languages.

Statistics: As per the 2011 Census report on Indian languages [6], India has 22 major or scheduled languages, each with a substantial volume of written content. Our dataset comprises **ten** languages — Kannada, Odia, Punjabi, Hindi, Gujarati, Marathi, Malayalam, Telugu, Tamil, and Bengali - representing diverse regions of India. Languages not included from the list of 22 major languages either share scripts with the included languages [1] or have minimal usage in the context of scene text in natural Indian settings [2]. Therefore, gathering data samples for these languages is challenging, as they may not be prevalent daily.

[1] Bodo, Dogri, Kashmiri, Konkani, Maithili, Nepali, and Sindhi
[2] Santali, Sanskrit

The dataset comprises ten languages, each with 25,000 word-level images. Among them, 17,500 and 2,500 word images are for training and validation, while 5,000 are reserved for testing.

Characteristics: Our STR dataset comprises regular and irregular samples. Regular datasets, exemplified by MLT-17 [17] and MLT-19 [16], mainly consist of frontal, horizontal word-level images with a small portion of distorted samples. In contrast, irregular datasets feature perspective text, low-resolution, and multi-oriented word images, posing challenges for STR. Our dataset, extracted from Indian road scenes, offers a diverse range of irregular samples (Fig. 4), reflecting the varied text instances encountered in real-world scenarios. By encompassing such a broad spectrum of linguistic and visual characteristics, the dataset serves as a valuable resource for advancing research in this field and enhancing the accuracy and versatility of scene text recognition systems.

4 Word Image Recognition from Indic Scene Images Challenge

To achieve accuracy in scene text recognition across diverse languages, this competition focused on nearly all Indian languages, which together represent about 17% of the world's population and share syntactic and semantic similarities.

4.1 Task

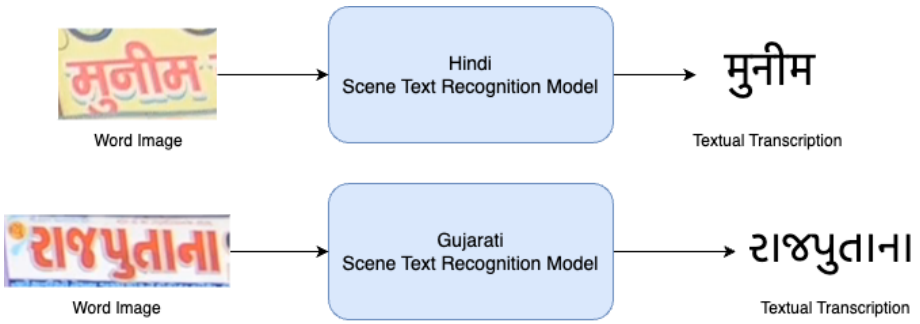

Fig. 5: Competition Task: given Indic scene text word-level images, the task is to recognize them.

The challenge comprised only one task (Fig. 5) - cropped word image recognition. The participants had to predict the text in all the cropped word images of the test set in ten targeted Indian languages.

Pre-Training Due to the scarcity of annotated real datasets, it is common practice to pre-train STR models on large synthetically generated datasets. For example, [10] introduced the MJSynth and [9] the SynthText dataset. While Latin STR, especially English, has moved to training exclusively on real datasets [1], the situation is different for Indian languages. Studies like [14], [8], and [13] have relied on synthetic datasets with over a million samples, followed by fine-tuning on smaller real datasets. However, traditional methods of combining text and background through image editing, as used in these works, fall short of accurately replicating real data distributions. Participants were encouraged to either use established synthetic datasets like [13] or explore newer approaches by [21], [20], and [22] for this competition.

4.2 Evaluation Metrics

In scene text recognition, the predicted text string is directly compared to the ground truth. Performance is evaluated at the character level by counting correctly recognized characters, and at the word level by checking if the predicted word exactly matches the ground truth.

Recognizing the importance of both metrics for assessing different aspects of the model's performance, the evaluation for this task included the Word Recognition Rate (WRR) at the word level and the Character Recognition Rate (CRR) at the character level. A prediction was deemed correct only if all characters matched exactly at every position.

$$WRR = \frac{W_r}{W} \times 100\%, \tag{1}$$

where W is the total number of words, and W_r represents the number of correctly recognized words.

At the character level, the Character Recognition Rate (CRR) represents the percentage of correctly recognized characters out of the total number of characters in a dataset. A higher CRR indicates better performance of the recognition model.

The CRR can be mathematically expressed as:

$$CRR = \frac{N_c}{N_t} \times 100\%, \tag{2}$$

where N_c is the number of correctly recognized characters and N_t is the total number of characters.

We assessed models and ranked them based on WRR and CRR individually, awarding them accordingly for each metric.

4.3 Submitted Methods

Participants could submit up to 10 entries, with their best-performing submission considered as the official entry for the competition. In total, 10 submissions were received from 5 teams. Table 2 presents the teams and the results of their top

submissions. Below is a summary of the top submissions from each team, along with a description of the baseline established when releasing the test set.

Baseline — As our baseline, we employed the CRNN model proposed by [7]. The network is composed of four key modules: a Transformation Network (TN), a Feature Extractor (FE), Sequence Modeling (SM), and Predictive Modeling (PM). The TN includes six plain convolutional layers, while the FE module is built on a ResNet architecture. The SM module features a two-layer Bidirectional LSTM (BLSTM) with 256 hidden units per layer. Finally, the PM module uses Connectionist Temporal Classification (CTC) to decode and recognize characters by aligning the feature sequence with the target character sequence.

Table 2: WRR and CRR values for the top submitted results by participating teams, alongside the baseline we provided. The highest CRR and WRR values for each language are highlighted in green. The results encompass all ten targeted languages.

Team Name	**Bengali**		**Kannada**		**Hindi**		**Telugu**		**Gujarati**	
	CRR	WRR	CRR	WRR	CRR	WRR	CRR	WRR	CRR	WRR
MVu	95.88%	88.22%	97.34%	92.4%	95.11%	89.68%	95.07%	85.92%	67.88%	46.76%
Alias	92.49%	81.76%	95.54%	87.04%	95.85%	88.20%	91.78%	80.00%	–	–
Baseline	94.04%	82.38%	96.24%	87.44%	96.48%	89.18%	94.03%	80.72%	64.02%	34.24%
TSNUK	90.46%	79.54%	94.84%	86.72%	94.35%	86.70%	90.56%	77.68%	67.67%	45.46%
KSK	85.36%	68.44%	79.78%	55.52%	88.52%	76.36%	69.49%	37.06%	61.51%	33.82%
Visionary Minds	82.91%	57.08%	76.57%	41.26%	75.67%	47.84%	77.68%	50.0%	61.49%	30.24%

Team Name	**Malayalam**		**Marathi**		**Odia**		**Punjabi**		**Tamil**	
	CRR	WRR	CRR	WRR	CRR	WRR	CRR	WRR	CRR	WRR
MVu	97.7%	93.18%	96.69%	92.58%	95.32%	86.58%	94.05%	85.06%	93.46%	79.7%
Alias	–	–	97.47%	92.62%	94.09%	86.80%	93.99%	85.06%	94.48%	81.50%
Baseline	97.18%	89.46%	97.06%	91.20%	94.62%	84.38%	93.84%	82.48%	94.82%	80.44%
TSNUK	95.85%	88.22%	95.60%	87.98%	91.21%	79.20%	91.49%	79.92%	90.38%	73.88%
KSK	70.52%	38.48%	91.04%	79.84%	82.41%	61.78%	90.43%	78.02%	66.83%	34.82%
Visionary Minds	–	–	88.12%	68.42%	85.45%	63.9%	85.66%	66.72%	81.22%	49.9%

MVu — This team utilized the PARSeq [2] architecture to develop a single model capable of handling multiple languages. A language token identifier (*language_token_id*) was integrated into the training process to achieve this. This token was added after the Beginning of Document (BOD) token and before the ground truth transcription, helping the model to identify and transcribe the language accurately. Due to the limited training data, the model was pretrained on synthetic data provided by [13] to improve its performance across all languages. This approach aimed to create a unified, robust solution for multilingual text recognition.

Alias — This team also utilized the PARSeq [2] model, training separate instances for each targeted Indian language using the provided real data without pre-training on synthetic or other real data. They made two notable modifications: reducing the batch size from 365 to 64 and turning off mixed precision training. Additionally, no data augmentation techniques were applied.

TSNUK — The participant used a variant of the traditional Convolutional Recurrent Neural Network (CRNN) [18], substituting Bidirectional GRU RNNs for BiLSTMs. Input images were resized to a fixed height while preserving aspect ratios and then processed through a CNN to extract feature maps. These feature maps were fed into a Bidirectional GRU RNN to capture sequential dependencies. The RNN's output was processed by a transcription layer using Connectionist Temporal Classification (CTC) loss for training. A token-passing decoding algorithm was employed to identify the most likely label sequence for recognition. Additionally, pre-training on synthetic data or data augmentation was not applied.

KSK — This participant's method utilizes a Swin Transformer as the encoder to extract visual features from images. These features are then decoded into text sequences using GPT-2, a pre-trained language model. Labels are tokenized with GPT2Tokenizer, set to a maximum length of 128 tokens. The model is trained using Hugging Face's Seq2SeqTrainer for 30 epochs with a batch size of 32.

Visionary Minds — This team employs a CRNN with unidirectional LSTM layers for sequence prediction, incorporating image augmentation and normalization during training. They utilize the Connectionist Temporal Classification (CTC) loss function to manage sequence prediction tasks with unknown input-output alignments effectively.

4.4 Result

Following the process described in Section 2, points were awarded for each team's final submissions across all languages, and the total scores (as shown in Table 3) were used to determine the competition's winner and runner-up. *MVu* emerged as the winner by excelling in nearly all languages across both metrics, while team *Alias*, with the best results in three languages and second-best in the others, secured the runner-up position (see Table 2). Sample text predictions from submitted results on the competition dataset are shown in Fig. 6.

4.5 Analysis

Three of the five final submissions used transformer or attention-based character decoding, while two employed CRNN models with CTC-based decoding.

[3] Bengali, Kannada, Hindi, Telugu, Gujarati, Malayalam, Marathi, Odia, Punjabi, and Tamil—arranged in order from top to bottom and left to right

Fig. 6: Example word images from 10 different languages[3], each accompanied by the ground truth text centered above the image. Below each word image, the predicted text from the *MVu* team is shown at the bottom left, and the predicted text from the *TSNUK* team is shown at the bottom right. Errors in the predictions are highlighted in red, with missing characters indicated by red boxes.

Table 3: **Final Scores and Language Coverage:** This table presents the final scores and the number of languages for which each team submitted transcriptions. The teams are listed in ranking order, with *MVu* as the winner and *Alias* as the runner-up.

Team Name	**# Languages**	**Total Points** (WRR)	**Total Points** (CRR)
MVu	10	47	47
Alias	8	36	35
TSNUK	10	32	32
KSK	10	14	14
Visionary Minds	9	9	9

Performance varied significantly across languages, with an average WRR standard deviation exceeding 10%, calculated from ten languages per team and averaged across all teams. For instance, despite their script similarities, the winning model exhibited a 40% WRR difference between Hindi and Gujarati, and a 13.5% difference between Tamil and Malayalam, which have some visual commonalities. Both *MVu* and *Alias* used PARSeq models, but *MVu* applied a single model across multiple scripts, while *Alias* used one model per script. The multi-script model, pre-trained on synthetic data, slightly outperformed the mono-script model trained only on real data for most languages, though it was not conclusively superior. CRNNs showed competitive results. The baseline CRNN model achieved performance comparable to *Alias's* single-script PARSeq model using only the real dataset. Replacing BiLSTM with BiGRU led to a slight performance drop, likely due to LSTM's better ability to capture long-term dependencies, which is significant given the long average word length in Indian languages. The performance decrease was more pronounced without bidirectionality, highlighting the benefit of processing sequences in both directions. Finally, the Swin Transformer combined with GPT-2 did not demonstrate a clear performance advantage over traditional transformer models.

5 Conclusions and Future Direction

This paper has summarized the WIRIndic 2024 competition, which was conducted to advance scene text recognition solutions for 10 Indian languages. The competition introduced a first-of-its-kind dataset for the Indic space, comparable in scale and diversity to existing Latin and other multilingual datasets, offering a significant opportunity for improving scene text solutions for Indian languages. The competition focused solely on word image recognition and was hosted on

the web portal https://ilocr.iiit.ac.in/icpr_2024_wirindic/, attracting 49 registrations and five team submissions. The competition provided real-time result updates throughout the event and will continue to publish results from post-competition submissions on the released datasets.

The submitted solutions showcased strong performance with several approaches, including a mix of attention-based and CTC-based methods. The top-performing solution, which employed a single multilingual model for all 10 languages, was particularly promising, highlighting the potential of leveraging linguistic commonalities to train larger, more effective models that can address multiple languages simultaneously. Although the winning solution achieved an average WRR score of 84.01% across the 10 languages, surpassing the baseline by a good margin, this result suggests that we are only beginning to develop serious, well-performing solutions for Indic STR. There remains significant room for improvement, especially in languages like Gujarati, Tamil, and Telugu. The performance variations among languages indicate that more work is needed to fully harness the potential of the dataset released through this competition.

Looking ahead, we are optimistic about the future of scene text recognition. Continued efforts to refine and build upon the solutions proposed in this competition, coupled with the now available rich dataset, will pave the way for more robust and accurate models. We hope this competition serves as a foundation for future research and development, leading to more generalized scene text solutions.

Acknowledgment

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

References

1. Baek, J., Matsui, Y., Aizawa, K.: What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
2. Bautista, D., Atienza, R.: Scene text recognition with permuted autoregressive sequence models. In: European conference on computer vision (2022)
3. Chen, X., Jin, L., Zhu, Y., Luo, C., Wang, T.: Text recognition in the wild: A survey. ACM Computing Surveys (CSUR) (2021)
4. Chng, C.K., Liu, Y., Sun, Y., Ng, C.C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., Liu, J., Karatzas, D., Chan, C.S., Jin, L.: Icdar2019 robust reading challenge on arbitrary-shaped text - rrc-art. In: 2019 International Conference on Document Analysis and Recognition (ICDAR) (2019)
5. Garcia-Bordils, S., Mafla, A., Biten, A.F., Nuriel, O., Aberdam, A., Mazor, S., Litman, R., Karatzas, D.: Out-of-vocabulary challenge report. In: European Conference on Computer Vision (2022)
6. GOI: Government indian language report. https://censusindia.gov.in/census.website/ (2011)

7. Gongidi, S., Jawahar, C.: iiit-indic-hw-words: A dataset for indic handwritten text recognition. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part IV 16 (2021)
8. Gunna, S., Saluja, R., Jawahar, C.: Transfer learning for scene text recognition in indian languages. In: Document Analysis and Recognition–ICDAR 2021 Workshops: Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16 (2021)
9. Gupta, A., Vedaldi, A., Zisserman, A.: Synthetic data for text localisation in natural images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016)
10. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Synthetic data and artificial neural networks for natural scene text recognition. arXiv preprint arXiv:1406.2227 (2014)
11. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Towards end-to-end unified scene text detection and layout analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
12. Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., Raptis, M.: Icdar 2023 competition on hierarchical text detection and recognition. In: International Conference on Document Analysis and Recognition (2023)
13. Lunia, H., Mondal, A., Jawahar, C.: Indicstr12: A dataset for indic scene text recognition. In: International Conference on Document Analysis and Recognition (2023)
14. Mathew, M., Jain, M., Jawahar, C.: Benchmarking scene text recognition in devanagari, telugu and malayalam. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (2017)
15. Mathew, M., Singh, A.K., Jawahar, C.: Multilingual ocr for indic scripts. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS) (2016)
16. Nayef, N., Patel, Y., Busta, M., Chowdhury, P.N., Karatzas, D., Khlif, W., Matas, J., Pal, U., Burie, J.C., Liu, C.l., Ogier, J.M.: Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition — rrc-mlt-2019. In: 2019 International Conference on Document Analysis and Recognition (ICDAR) (2019)
17. Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., Khlif, W., Luqman, M.M., Burie, J.C., Liu, C.l., Ogier, J.M.: Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification - rrc-mlt. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (2017)
18. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. PAMI (2016)
19. Tom, G., Mathew, M., Garcia-Bordils, S., Karatzas, D., Jawahar, C.: Icdar 2023 competition on roadtext video text detection, tracking and recognition. In: Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part II (2023)
20. Wu, L., Zhang, C., Liu, J., Han, J., Liu, J., Ding, E., Bai, X.: Editing text in the wild. In: Proceedings of the 27th ACM international conference on multimedia (2019)
21. Xie, Y., Chen, X., Zhan, H., Shivakumara, P., Yin, B., Liu, C., Lu, Y.: Weakly supervised scene text generation for low-resource languages. Expert Systems with Applications (2024)
22. Yang, Q., Huang, J., Lin, W.: Swaptext: Image based texts transfer in scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)