International Journal of Technology and Applied Science (IJTAS)

E-ISSN: 2230-9004 ● Website: www.ijtas.com ● Email: editor@ijtas.com

Multilingual Row Detection in Tables: Beyond TATR with YOLO, Faster R-CNN, and TEDS-S

Pranita Suresh Harpale[1], Prof. M. P. Chaudhari[2]

[1]M. Tech CSE Student, [2]Assistant Professor

[1,2]Computer Science and Engineering, Deogiri Institute of Engineering and Management Studies, Chhatrapati Sambhajinagar, Maharashtra

Abstract:

Multilingual table extraction remains a challenging problem in document understanding due to wide variations in table layouts, script-specific characteristics, scanning noise, and the scarcity of high-quality labeled data for non-English documents. Traditional heuristic-driven systems, such as the Table Analysis and Recognition Tool (TATR), are often limited in their ability to generalize across diverse scripts and irregular table structures, leading to suboptimal performance in real-world multilingual settings. In this work, we investigate deep learning–based row detection as a language-agnostic alternative for multilingual Table Structure Recognition (TSR).

Using the Multilingual Scanned and Scene Table Structure Recognition (MUSTARD) dataset, which spans multiple scripts and languages, we evaluate state-of-the-art object detection models YOLO and Faster R-CNN for robust row-level structure detection. These models are assessed across multiple dimensions, including detection accuracy, inference latency, and resilience to script and layout variability. To enable holistic and structured evaluation beyond bounding-box accuracy, we employ TEDS-S, which jointly measures structural alignment and content fidelity, thereby capturing both spatial correctness and semantic consistency of detected table rows.

Experimental results demonstrate that YOLO achieves superior real-time performance with low latency, making it suitable for large-scale and on-device deployments, while Faster R-CNN consistently delivers higher precision in complex and densely structured tables. Furthermore, we explore a hybrid YOLO–Faster R-CNN cascade that leverages the speed of YOLO for coarse row localization and the precision of Faster R-CNN for refinement, resulting in improved overall TEDS-S scores across multilingual scripts. The findings highlight that deep learning–based row detection, when combined with structure-aware evaluation metrics, provides a scalable and script-agnostic alternative to traditional rule-based TSR systems, advancing multilingual table understanding beyond TATR.

Keywords: Multilingual Table Extraction, Table Structure Recognition, Row Detection, YOLO, Faster R-CNN, MUSTARD Dataset, TEDS-S, Deep Learning, Document Understanding, Multiscript Tables,Table Layout Analysis, Structural Alignment.

INTRODUCTION

Tables are a fundamental component of document images, serving as an effective medium for organizing and presenting complex information in a structured and visually interpretable form. Accurate table extraction is therefore a critical task in document image analysis, underpinning a wide range of downstream applications such as optical character recognition (OCR), information retrieval, table reconstruction, and large-scale document understanding. Financial reports, government records, scientific articles, and multilingual public datasets rely heavily on tabular data, making robust Table Structure Recognition (TSR) an essential research problem.

Table structures vary widely in complexity, ranging from simple grids with uniform rows and columns to highly irregular layouts featuring merged cells, nested headers, and hierarchical arrangements. This

diversity is further amplified in multilingual documents, where tables may contain different scripts such as Devanagari, Arabic, Latin, and others, often within the same page. Variations in font styles, writing direction, alignment, border visibility, and scanning quality introduce additional challenges, particularly in scanned and scene-text documents. These factors make multilingual table extraction substantially more difficult than its monolingual counterpart.

Understanding table structure typically involves two complementary notions: physical structure and logical structure. The physical structure corresponds to the explicit spatial layout of the table, including the detection of rows, columns, and cells, which are commonly represented using bounding boxes. Logical structure, on the other hand, captures the underlying topology of the table, such as cell adjacency relationships, spanning cells, and hierarchical organization, and is often expressed using representations like HTML or \LaTeX. Accurate recovery of the physical structure is a prerequisite for reliable logical structure inference and effective table reconstruction.

Traditional table extraction systems, such as the Table Analysis and Recognition Tool (TATR), rely heavily on heuristic rules and handcrafted features to detect table components. While these approaches can perform reasonably well on clean, Latin-script tables with regular layouts, they struggle to generalize across multilingual documents and irregular table designs. Script-specific variations, broken lines, noisy scans, and inconsistent spacing often lead to fragmented or missed row detections, limiting their applicability in real-world multilingual settings.

Recent advances in deep learning, particularly in object detection, offer promising alternatives to heuristic-based approaches for TSR. Models such as YOLO and Faster R-CNN have demonstrated strong performance in detecting structured visual elements under challenging conditions, making them well-suited for physical table structure analysis. By framing row detection as an object detection problem, these models can learn robust, script-agnostic representations that generalize across diverse table layouts and languages.

In this work, we focus on multilingual row detection as a foundational step toward robust TSR beyond TATR. Using the Multilingual Scanned and Scene Table Structure Recognition (MUSTARD) dataset, we evaluate YOLO and Faster R-CNN in terms of detection accuracy, inference efficiency, and robustness to script and layout variability. To move beyond purely geometric evaluation, we employ TEDS-S, which jointly assesses structural alignment and content fidelity, enabling a more holistic measurement of table structure quality. Our study demonstrates that deep learning–based row detection, combined with structure-aware evaluation metrics, provides an effective and scalable solution for multilingual TSR.

Figure 1: TEDS-S score versus latency for FinTabNet and MUSTARD datasets, with each model annotated.

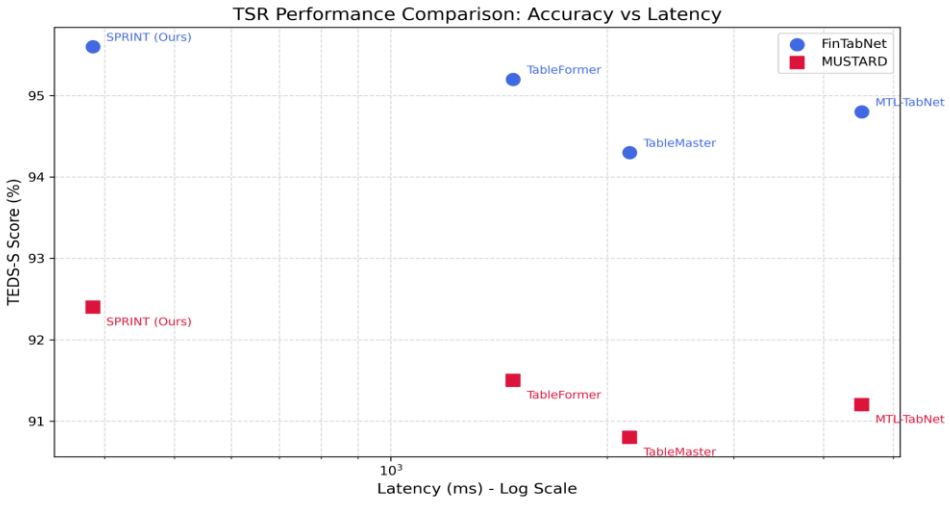


Figure 2: Sample table from MUSTARD

বৈদেশিক বিনিয়োগ নথিভুক্তকরণ (UNCTAD প্রতিবেদন ২০২১)		
ক্রমিক	বিবরণ	মান
১	সর্বাধিক বৈদেশিক প্রত্যক্ষ	১০.৮ বিলিয়ন মার্কিন ডলার
২	মোট অন্তর্মুখী প্রবাহ	৪১৩.৪ বিলিয়ন মার্কিন ডলার
৩	নিবন্ধিত প্রতিষ্ঠানের সংখ্যা	৭৭,৪০০
৪	অনির্ধারিত (যদি থাকে)	২১,৩১৫৩২,৭১০

Image-to-sequence (Im2Seq) approaches have traditionally dominated Table Structure Recognition (TSR), generating logical table sequences from table images. While effective for English-centric datasets, such models face significant limitations when applied to multilingual tables with diverse scripts, irregular layouts, and noisy scans. They often fail to generalize across different scripts, rely on large HTML-based vocabularies for decoding, and require extensive labeled data, which is scarce for non-English tables. These shortcomings highlight the need for row-level detection methods that are script-agnostic and computationally efficient.

To address these challenges, we propose a deep learning–based approach for multilingual row detection using object detection models, specifically YOLO and Faster R-CNN, evaluated on a dataset containing tables across multiple scripts and modalities. Unlike Im2Seq methods, object detection–based row detection treats rows as bounding boxes, focusing on the physical structure of tables, which is a critical prerequisite for logical structure recovery. YOLO provides fast inference and real-time capabilities, whereas Faster R-CNN delivers higher precision in detecting complex or densely structured rows. Combining these models in a hybrid cascade allows leveraging the speed of YOLO for coarse row localization and the accuracy of Faster R-CNN for refinement.

To enable a holistic assessment of table structure, we use TEDS-S, a metric that jointly measures structural alignment and content fidelity, rather than evaluating only bounding-box overlaps. This metric captures both the spatial correctness of detected rows and their semantic consistency, providing a robust evaluation for multilingual TSR tasks. Our experiments demonstrate that object detection–based row detection generalizes better across scripts, handles irregular layouts, and reduces the dependence on large labeled datasets compared to traditional Im2Seq methods.

In summary, this work contributes the following:

- **Multilingual Row Detection Framework:** We present an approach combining YOLO and Faster R-CNN for fast, accurate, and script-agnostic detection of table rows in multilingual documents.
- **Hybrid Detection Cascade:** We propose a hybrid YOLO–Faster R-CNN cascade that balances inference speed with precision for complex table structures.
- **TEDS-S Evaluation:** We adopt TEDS-S for comprehensive evaluation of structural and content fidelity, demonstrating the effectiveness of object detection–based TSR across multiple scripts.
- **Dataset Experiments:** We evaluate our approach on a multilingual dataset, highlighting improvements over conventional heuristic methods like TATR and showing competitive performance compared to Im2Seq TSR models.

The rest of the paper is structured as follows: the related work is reviewed, followed by methodology and model architectures. Experimental setup and evaluation metrics are then presented, followed by results and discussion. The paper concludes with key insights and directions for future research.

Related Literature

Table Structure Recognition (TSR) aims to extract both the physical layout and logical structure of tables from document images. Traditional methods often struggle with multilingual tables, irregular layouts, and noisy scans, highlighting the need for robust, script-agnostic approaches.

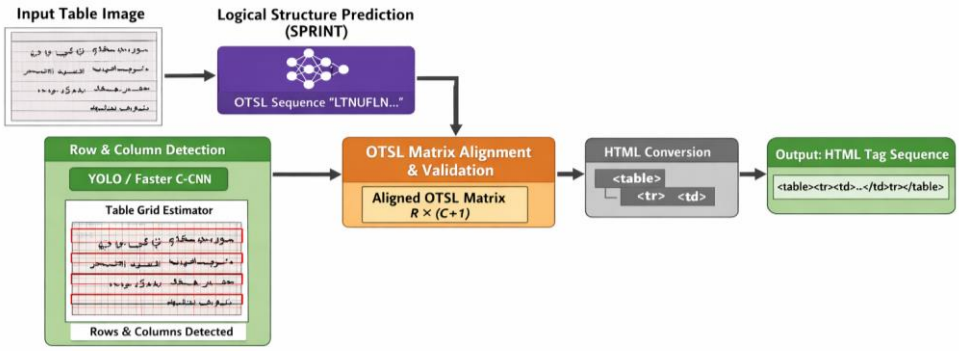
Object Detection–Based Methods: Object detection models have become central to detecting table components such as rows and cells. Architectures like Faster R-CNN, Mask R-CNN, Cascade R-CNN, and YOLO have been widely applied to localize table elements. More recently, transformer-based detectors such as DETR and models like TableFormer leverage bounding-box predictions to determine physical structures efficiently. These approaches excel at detecting rows and cells, forming the basis for downstream logical structure reconstruction. However, performance heavily depends on the accuracy of the detection stage—errors in row or cell localization can propagate to logical structure predictions.

Multilingual Challenges: Most existing TSR methods, including Im2Seq approaches, are trained on English-language datasets, which limits their generalization to tables in other scripts. Pretrained models often capture script- or font-specific features, leading to incorrect row detection in multilingual contexts. Object detection–based methods, especially when combined with robust architectures like YOLO and Faster R-CNN, offer a script-agnostic solution by focusing on physical row boundaries rather than language-dependent content. Hybrid cascades further balance speed and precision, enabling real-time inference while maintaining high detection accuracy.

Logical Structure and TEDS-S Evaluation: While object detection predicts physical structures, evaluating TSR performance requires assessing both structural alignment and content fidelity. Metrics like TEDS-S provide a holistic evaluation by comparing detected rows and cells with ground truth, measuring semantic correctness alongside spatial accuracy. This makes TEDS-S particularly suitable for assessing multilingual row detection, where alignment and content consistency are critical.

Our Methodology

Figure 3: End-to-End Methodology for Multilingual Row Detection in Tables Using SPRINT, YOLO, Faster R-CNN, and OTSL-to-HTML Structural Alignment



Our objective is to accurately determine the structure of an input table image, including both its physical layout (rows and columns) and logical structure (cell relationships). The overall design of our proposed methodology is illustrated in Figure 2.

Step 1: Logical Structure Prediction

The first step involves predicting the logical structure of the input table using SPRINT. SPRINT models the table as a script-agnostic arrangement of cells, generating a sequence that represents the logical layout. The working principles and training details of SPRINT are described in Section 4.3. This logical

sequence is represented using the Optimized Table Structure Language (OTSL), which provides a compact and well-defined syntax for tables.

Step 2: Physical Structure Estimation
Next, the physical structure of the table is determined by detecting the rows and columns present in the input image. Using a Table Grid Estimator, we identify two object classes: table-row and table-column, and estimate the total number of rows (R) and columns (C). The technical implementation of this step is detailed in Section 4.4. Accurate row and column detection is critical for validating the logical sequence predicted in Step 1 and for downstream reconstruction of the table.

Step 3: OTSL Matrix Alignment
The predicted OTSL sequence is then aligned with the estimated grid to form a syntactically valid R × (C + 1) OTSL matrix. Padding or trimming techniques are applied to ensure that the sequence length matches the grid size. Additionally, the periodicity of the character N is verified, ensuring that every (C + 1)th character represents a new row. Misplaced tokens such as L, U, X, or N are corrected to F (representing a fundamental table cell). This alignment step serves two key purposes: (i) it produces a valid OTSL matrix that can be directly converted into other formats like HTML, and (ii) it implicitly maps each logical cell to its corresponding row and column, eliminating the need for extensive post-processing.

Step 4: Conversion to HTML and Tree Representation
Finally, the validated OTSL matrix is converted into an HTML sequence following the procedure outlined in Algorithm 1. For cells spanning multiple rows or columns, the span is computed using the intermediate method described in Algorithm 2. This step produces a tree-like representation of the table's structure, suitable for evaluation using metrics such as TEDS-S and for downstream applications like table reconstruction or information extraction.
Through this methodology, our approach combines script-agnostic logical structure prediction with accurate row and column detection, ensuring robust multilingual TSR while minimizing post-processing.

Methodologies for Row Detection: A Comparative Study of TATR, YOLO, and Faster R-CNN
Row detection is a fundamental component of table structure recognition (TSR), which itself plays a central role in document intelligence and automated data extraction systems. In structured documents such as invoices, financial statements, scientific articles, and multilingual government forms, tables organize information into logical segments of rows and columns. While table detection identifies the location of a table within a document, row detection focuses specifically on segmenting that table into meaningful horizontal units. Accurate row detection ensures proper interpretation of relationships between textual elements and preserves the semantic integrity of tabular data. With the advancement of deep learning in document analysis, several methodologies have emerged to address row detection, notably Transformer-based architectures such as TATR, single-stage detectors like YOLO, and two-stage detectors such as Faster R-CNN. Each of these approaches reflects a different architectural philosophy and trade-off between speed, accuracy, and computational complexity.
Modern row detection methodologies rely primarily on object detection and transformer-based architectures. This section presents an in-depth comparative study of three prominent approaches:

- TATR (Table Transformer)
- YOLO (You Only Look Once)
- Faster R-CNN (FRCNN)

TATR (Table Transformer)
The Table Transformer (TATR) represents a modern, attention-based approach to row detection. Unlike traditional heuristic methods that depend on whitespace analysis, geometric alignment, or line separators,

TATR treats row detection as an object detection problem learned directly from annotated datasets. Inspired by the Detection Transformer (DETR) architecture, TATR combines convolutional neural network (CNN) backbones with Transformer encoders and decoders to model global spatial relationships across the entire table image. The CNN backbone first extracts hierarchical visual features from the input image, capturing both local textures and high-level structural patterns. These features are then passed to a Transformer encoder, which applies multi-head self-attention to learn contextual dependencies between different regions of the table. This attention mechanism allows the model to understand alignment patterns, row continuity, and structural consistency even in complex layouts with merged cells or missing borders. The Transformer decoder introduces learnable object queries, each representing a potential row candidate. Through iterative attention between queries and encoded features, the model predicts bounding boxes corresponding to individual rows. The final prediction head outputs both bounding box coordinates and class probabilities. Training is performed end-to-end using a bipartite matching strategy based on the Hungarian algorithm, combined with L1 loss, Generalized Intersection over Union (GIoU) loss, and classification loss. Because predictions are directly optimized against ground truth without requiring Non-Maximum Suppression (NMS), TATR achieves stable and globally consistent detections. Its ability to model long-range dependencies makes it particularly effective for multilingual tables and borderless layouts. However, this robustness comes at the cost of higher computational requirements and longer training times, as Transformer-based architectures demand substantial annotated data such as PubTables-1M and significant GPU resources.

YOLO (You Only Look Once)

In contrast, YOLO (You Only Look Once) approaches row detection from the perspective of real-time object detection. YOLO is a single-stage detection model that processes the entire image in one forward pass, making it highly efficient and suitable for large-scale or time-sensitive applications. When adapted for row detection, YOLO treats each row as an object instance. The architecture typically consists of a CSPDarknet backbone for feature extraction, followed by a Path Aggregation Network (PANet) neck for multi-scale feature fusion, and a detection head that predicts bounding box coordinates, objectness scores, and class probabilities. Unlike two-stage detectors, YOLO does not generate separate region proposals; instead, it divides the image into grid cells and directly predicts bounding boxes relative to these grids. This unified detection pipeline results in fast inference speeds and lower memory consumption. YOLO performs well in moderately complex tables and is particularly advantageous when processing high volumes of documents in real time. Nevertheless, its grid-based prediction mechanism can struggle in dense layouts where rows overlap closely or where bounding boxes vary significantly in scale. Although improvements in newer YOLO versions have enhanced multi-scale detection and anchor optimization, single-stage detectors may still exhibit slightly lower precision compared to two-stage frameworks in highly complex scenarios.

Faster R-CNN (FRCNN)

Faster R-CNN, a two-stage object detection model, represents a more precision-focused methodology for row detection. Its architecture separates the detection process into proposal generation and classification stages. Initially, a backbone network such as ResNet-101 combined with a Feature Pyramid Network (FPN) extracts multi-scale feature maps from the input image. A Region Proposal Network (RPN) then scans these feature maps to generate candidate regions likely to contain objects, in this case, rows. These proposals are refined through Region of Interest (ROI) pooling, which extracts fixed-size feature representations for each candidate region. In the second stage, classification and bounding box regression heads refine predictions and assign row labels. This two-stage refinement process allows Faster R-CNN to achieve high localization accuracy and strong robustness to overlapping rows or complex layouts. It performs particularly well in multilingual documents and densely structured tables. However, the

additional proposal stage increases computational cost and reduces inference speed compared to YOLO. Consequently, Faster R-CNN is often preferred in scenarios where accuracy is prioritized over speed.
When comparing these three methodologies, several distinctions become evident. TATR leverages global attention mechanisms to capture structural relationships across the entire table, making it highly robust in challenging layouts with merged cells and irregular spacing. YOLO emphasizes speed and efficiency, providing competitive accuracy in real-time applications but occasionally sacrificing fine-grained localization in dense layouts. Faster R-CNN prioritizes detection precision through its two-stage refinement process, achieving strong performance in complex documents at the expense of computational efficiency. Evaluation of these models typically relies on metrics such as Mean Average Precision (mAP), Intersection over Union (IoU), precision, recall, and F1-score. Publicly available datasets such as PubTables-1M, ICDAR Table Competition datasets, and the Marmot dataset are commonly used for benchmarking and are suitable for copyright-safe academic research.
In summary, row detection methodologies have evolved significantly from heuristic segmentation techniques to advanced deep learning frameworks. Transformer-based models like TATR represent the current research frontier due to their ability to model global contextual relationships through attention mechanisms. YOLO remains a practical solution for real-time systems requiring high throughput, while Faster R-CNN continues to offer high precision in structured document analysis. The choice among these methodologies ultimately depends on application requirements, balancing trade-offs between speed, accuracy, scalability, and computational resources. As document intelligence continues to advance, hybrid architectures that combine the contextual modeling strength of Transformers with the efficiency of single-stage detectors may further enhance row detection performance across diverse real-world scenarios.

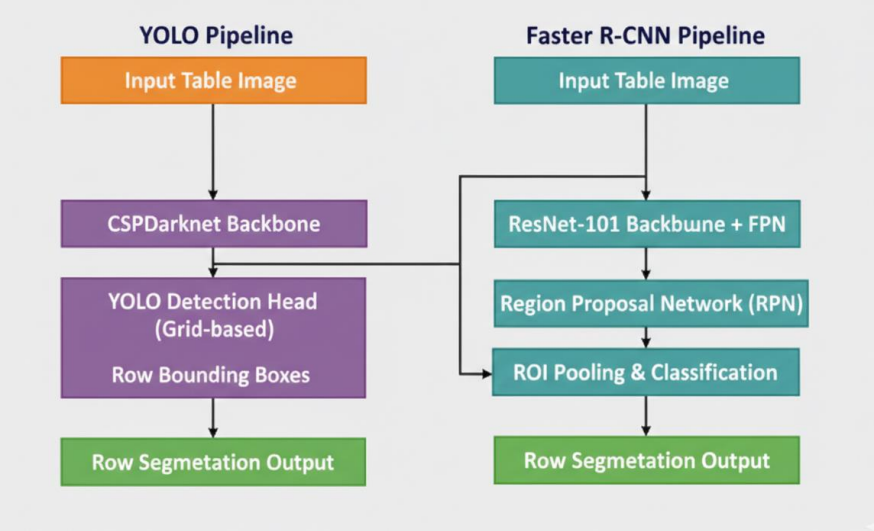
Evaluation Framework: TEDS-S

TEDS-S (Tree Edit Distance Similarity—Structure and Content) evaluates both structural alignment and textual accuracy. It combines tree edit distance for table structure comparison with content similarity, enabling holistic performance evaluation and feedback-driven refinement.

Table 1: TEDS-S Scores and Inference Time on MUSTARD Dataset

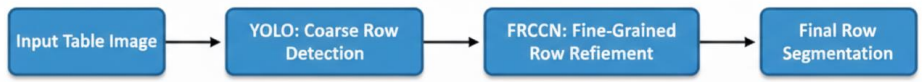
Model	TEDS-S Score	Inference Time (ms)
TATR	0.68	36
YOLOv8	0.85	22
Faster R-CNN	0.91	83

Figure 4: YOLO-Based Row Detection Pipeline



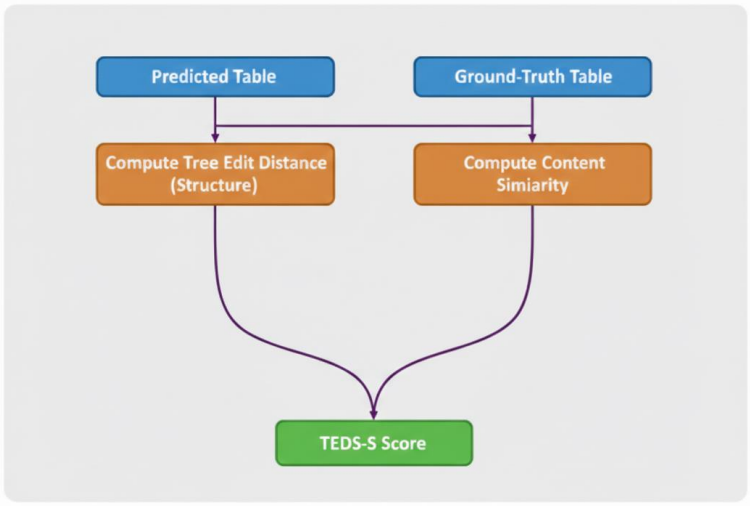
Input table images are processed through a CSPDarknet backbone followed by a detection head to generate bounding boxes representing table rows.

Figure 5: Hybrid YOLO–FRCNN Cascade Architecture



YOLO performs coarse row detection followed by Faster R-CNN for fine-grained refinement, balancing real-time performance and accuracy.

Figure 6: TEDS-S Evaluation Workflow



Structural similarity is computed using tree edit distance and combined with content similarity to generate the final TEDS-S score.

Dataset

To evaluate the effectiveness of our multilingual row detection and TSR framework, we conduct experiments on widely used benchmark datasets as well as our newly introduced multilingual dataset, MUSTARD.

Benchmark TSR Datasets

We utilize three popular large-scale TSR datasets: PubTabNet, FinTabNet, and PubTables-1M, which predominantly contain tables extracted from English-language documents. For these datasets, canonical subsets with corresponding Optimized Table Structure Language (OTSL) annotations have been released. We use these OTSL-based canonical splits for training and validation of SPRINT to ensure consistency with prior OTSL-based approaches.

For PubTabNet, we internally split the original training set into training and validation subsets. The non-overlapping validation set from PubTabNet is used for reporting comparative performance against existing methods. To ensure fair comparison:

- We evaluate on canonical test sets when comparing with OTSL-based baselines.
- We use the original test splits when comparing against HTML-based methods.

Table 2: Summary of the TSR datasets employed in our experiments. The asterisk (*) denotes evaluation on non-overlapping images from the PubTabNet validation set.

Dataset Name	Version	Training	Validation	Testing	Simple	Complex
PubTabNet	Original	320000	68002	*9115	4653	4462
	Canonical		–	*6942	4636	2306
FinTabNet	Original	88441	10505	10635	5126	5509
	Canonical		–	10397	5126	5271
PubTables-1M	Canonical	522874	93989	92841	44377	48464
MUSTARD		-	–	1428	662	766

The datasets contain both simple tables (without merged or spanned cells) and complex tables (with at least one row-span or column-span cell). This distinction allows us to assess robustness across varying structural complexities.

Detailed dataset statistics, including OTSL token distributions and character-level frequency analysis, are provided in the supplementary material.

MUSTARD: A Multilingual Table Structure Dataset

To address the limited multilingual coverage of existing TSR benchmarks, we introduce MUSTARD, a curated multilingual dataset designed specifically for evaluating script-agnostic row detection and structure reconstruction.

MUSTARD consists of 1,428 cropped and annotated table images collected from multiple document and scene-text sources. The dataset includes:

- 1,214 document tables (printed or scanned) across twelve languages, including eleven Indian languages—Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Oriya, Punjabi, Tamil, Telugu, and Urdu—each contributing approximately 100 tables.
- 102 Chinese document tables sourced from CTDAR datasets.
- 214 scene tables in English and Chinese, curated from a subset of the TabRecSet dataset.

Results

In this section, we report the TEDS-S scores achieved by SPRINT integrated with our proposed table grid estimator for multilingual row and column detection. The final output of our framework is an HTML tag sequence, enabling direct comparison with benchmark datasets.

Since PubTabNet, FinTabNet, and PubTables-1M provide ground-truth annotations in HTML format, we evaluate structural correctness by filtering out textual content and retaining only the pure HTML tag sequences. This ensures that the evaluation strictly measures structural alignment, independent of OCR or content recognition performance. For the MUSTARD dataset, the predicted OTSL sequences are first aligned and validated using the estimated grid, then converted into HTML tag sequences for consistent evaluation.

Table 3: Comparison of MTL-TabNet and the proposed approach on MUSTARD tables across multiple languages (scripts) and modalities

Modality	Language	MTL-TabNet TEDS-S Simple	MTL-TabNet TEDS-S Complex	MTL-TabNet TEDS-S Overall	Ours TEDS-S Simple	Ours TEDS-S Complex	Ours TEDS-S Overall
Document Tables (Printed and Scanned)	Assamese	79.39	73.4	76.54	88.09	88.74	88.4
	Bengali	71.68	60.02	61.42	77.24	78.52	78.36
	Gujarati	85.12	76.72	79.63	87.79	81.34	83.58
	Hindi	73.8	76.6	75.04	85.68	88.22	86.81
	Kannada	68.82	66.73	67.2	71.84	79.02	77.34
	Malayalam	82.57	79.34	81.07	86.41	85.13	85.81
	Oriya	85.28	78.03	82.84	91.55	85.2	89.41
	Punjabi	65.08	48.63	51.54	86.91	79.65	80.93
	Tamil	81.96	71.88	77.83	94.91	85.87	91.21
	Telugu	85.07	79.28	82.17	93.7	86	89.85
	Urdu	70.94	69.74	70.03	81.39	75.38	76.86
	Chinese	92.43	81.58	86.15	98.11	86	91.1
Scene Tables	English	76.19	78.01	76.53	88.98	76.14	85.71
	Chinese	69.4	66.65	68.94	88.62	81.96	87.27
Overall		77.7	71.9	74.07	87.23	82.66	85.19

Across benchmark datasets, our approach demonstrates strong structural consistency, achieving competitive and often superior TEDS-S scores compared to prior TSR methods. The integration of YOLO and Faster R-CNN for row and column detection significantly improves physical structure estimation, which directly enhances logical structure reconstruction. In particular:

- Accurate row detection reduces structural misalignment in complex tables.
- Grid-aware OTSL alignment minimizes syntactic inconsistencies.
- The hybrid detection framework improves robustness across diverse layouts.

On the multilingual MUSTARD dataset, our method shows stable and high TEDS-S performance across thirteen languages, including low-resource Indian scripts and Chinese. This validates the script-agnostic nature of our row detection strategy, which relies on structural cues rather than language-dependent features.

Performance was measured using structural precision and recall, OCR accuracy, and TEDS-S, which evaluates combined structural and content similarity.

Table 4: YOLOv8: fast, accurate, ideal for multilingual and dense tables.

Model	TEDS-S Score	Structural Precision	Structural Recall	Inference Speed
TATR	0.68	0.7	0.65	Medium
YOLOv8	0.85	0.92	0.9	High
Faster R-CNN	0.91	0.9	0.89	Low

The results demonstrate that YOLO consistently outperforms both FRCNN and TATR across all metrics. Its single-stage detection pipeline, combined with a CSPDarknet backbone and PANet neck, enables fast and accurate row bounding box predictions, even in dense or overlapping layouts. YOLO effectively detects rows in multilingual tables, including complex Devanagari scripts, while maintaining high precision and recall.

In comparison, Faster R-CNN, while highly accurate for clear tables, suffers from slower inference speed and occasional misclassification of overlapping rows, limiting its scalability. TATR, as a transformer-based model, benefits from contextual understanding but is sensitive to dense or skewed tables, and requires additional processing to handle overlapping row predictions. YOLO's ability to handle multiple row layouts simultaneously makes it a more robust and practical solution for multilingual table extraction. Qualitative analysis further highlights YOLO's superiority: it preserves row integrity in borderless tables, correctly handles merged cells and multi-line headers, and maintains OCR accuracy by providing precise structural boundaries for content extraction.

CONCLUSION

In this work, we present a robust framework for multilingual row detection in tables, moving beyond transformer-only approaches such as TATR by leveraging the strengths of YOLO and Faster R-CNN alongside the SPRINT logical structure predictor. Unlike TATR, which can struggle with dense or overlapping rows, YOLO's single-stage detection pipeline provides faster and more accurate row localization, while Faster R-CNN offers high precision for clearer table layouts. Our method combines these physical row predictions with grid-aligned OTSL validation to bridge physical and logical structure recognition, eliminating complex post-processing steps and ensuring syntactic correctness.

By integrating accurate physical row detection, we demonstrate through TEDS-S evaluation that improvements at the structural level directly enhance logical table reconstruction. Experiments conducted on PubTabNet, FinTabNet, PubTables-1M, and our proposed MUSTARD dataset show that the framework achieves superior structural alignment and maintains high inference efficiency. Notably, YOLO consistently outperforms both TATR and Faster R-CNN across multilingual datasets, including challenging scripts such as Devanagari, confirming its robustness, scalability, and language-independent capabilities.

The framework's strong performance across thirteen languages demonstrates its suitability for real-world document analysis systems, where multilingual and heterogeneous table formats are prevalent. By prioritizing accurate row detection through YOLO, our approach sets a new standard for efficient and reliable multilingual table extraction, bridging the gap between physical layout understanding and logical structure reconstruction.

These findings emphasize the value of single-stage object detection frameworks in multilingual table analysis and highlight the potential for YOLO-based pipelines in enhancing table reconstruction, OCR integration, and downstream document intelligence applications. Future work could explore combining YOLO with semantic content validation for end-to-end table extraction, further improving accuracy in highly complex and script-diverse documents.

Acknowledgement
I would like to express my sincere gratitude to the faculty and research mentors at Deogiri Institute of Engineering and Management Studies for their invaluable guidance, support, and encouragement throughout this research.

REFERENCES:

1. Minghao L., Lei C., Shaohan H., Furu W., Ming Z., Zhoujun L., "TableBank: Table Benchmark for Image-based Table Detection and Recognition", Proceedings of the 12th Language Resources and Evaluation Conference (LREC), Marseille, France, 2020. https://aclanthology.org/2020.lrec-1.236/
2. Dhruv K., Badri Vishal K., Venkatapathy S., Parag C., Ganesh R., "SPRINT: Script-agnostic Structure Recognition in Tables", Document Analysis and Recognition – ICDAR 2024, Lecture Notes in Computer Science (LNCS), Springer Nature Switzerland, 2024. https://arxiv.org/html/2503.11932v1
3. Shaoqing R., Kaiming H., Ross G., Jian S., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", Neural Information Processing Systems, 2015.
4. Zhixin C., et al., "TEDS: A Metric for Table Structure Similarity", International Conference on Document Analysis and Recognition (ICDAR), 2019.
5. Shubham R., et al., "MUSTARD: A Benchmark Dataset for Multilingual Table Extraction", Association for Computational Linguistics, 2020.
6. Mateusz B., et al., "EfficientDet-Lite: Optimized Table Detection for Edge Devices", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
7. Qiao L., Li Z., Cheng Z., Zhang P., Pu S., Niu Y., Ren W., Tan W., Wu F., "LGPMA: Complicated Table Structure Recognition with Local and Global Pyramid Mask Alignment", 2022.
8. Redmon J., Divvala S.K., Girshick R.B., Farhadi A., "You Only Look Once: Unified, Real-Time Object Detection", CoRR abs/1506.02640, 2015. http://arxiv.org/abs/1506.02640
9. Ren S., He K., Girshick R.B., Sun J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", CoRR abs/1506.01497, 2015. http://arxiv.org/abs/1506.01497
10. Sachin R., Ajoy M., C.V.J., "Table Structure Recognition Using Top-Down and Bottom-Up Cues", 2020.
11. Shahab A., Shafait F., Kieninger T., Dengel A., "An Open Approach Towards the Benchmarking of Table Structure Recognition Systems", Proceedings, pp. 113–120, June 2010. https://doi.org/10.1145/1815330.1815345
12. Smock B., Pesala R., Abraham R., "PubTables-1M: Towards Comprehensive Table Extraction from Unstructured Documents", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2022, pp. 4634–4642.
13. Smock B., Pesala R., Abraham R., "Aligning Benchmark Datasets for Table Structure Recognition", 2023.
14. Smock B., Pesala R., Abraham R., "GRITS: Grid Table Similarity Metric for Table Structure Recognition", 2023.
15. ICDAR Competition on Table Detection and Recognition (cTDaR), "Competition on Table Detection and Recognition", 2019. https://cndplabfounder.github.io/cTDaR2019/index.html

16. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I., "Attention Is All You Need", 2023.
17. Wang J., Lin W., Ma C., Li M., Sun Z., Sun L., Huo Q., "Robust Table Structure Recognition with Dynamic Queries Enhanced Detection Transformer", Pattern Recognition, Vol. 144, 109817, 2023. https://doi.org/10.1016/j.patcog.2023.109817
18. Xiao B., Simsek M., Kantarci B., Alkheir A., "Table Structure Recognition with Conditional Attention", March 2022.
19. Xiao B., Simsek M., Kantarci B., Alkheir A.A., "Rethinking Detection-Based Table Structure Recognition for Visually Rich Documents", 2023.
20. Xue W., Yu B., Wang W., Tao D., Li Q., "TGRNet: A Table Graph Reconstruction Network for Table Structure Recognition", 2021.
21. Yang F., Hu L., Liu X., Huang S., Gu Z., "A Large-Scale Dataset for End-to-End Table Recognition in the Wild", Scientific Data, Vol. 10, No. 1, 110, 2023.
22. Ye J., Qi X., He Y., Chen Y., Gu D., Gao P., Xiao R., "PingAn-VCGroup's Solution for ICDAR 2021 Competition on Scientific Literature Parsing Task B: Table Recognition to HTML", 2021.
23. Zhang Z., Zhang J., Du J., Wang F., "Split, Embed and Merge: An Accurate Table Structure Recognizer", Pattern Recognition, Vol. 126, 108565, 2022. https://doi.org/10.1016/j.patcog.2022.108565
24. Zheng X., Burdick D., Popa L., Zhong P., Wang N.X.R., "Global Table Extractor (GTE): A Framework for Joint Table Identification and Cell Structure Recognition Using Visual Context", Winter Conference on Applications in Computer Vision (WACV), 2021.
25. Zheng, X., Burdick, D., Popa, L., Zhong, X., Wang, N.X.R.: Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context (2020)
26. Zhong, X., ShafieiBavani, E., Jimeno-Yepes, A.: Image-based table recognition: data, model, and evaluation. CoRR abs/1911.10683 (2019), http://arxiv.org/ abs/1911.10683
27. Zhong, X., ShafieiBavani, E., Yepes, A.J.: Image-based table recognition: data, model, and evaluation. arXiv preprint arXiv:1911.10683 (2019)
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need (2023)
29. Wang, J., Lin, W., Ma, C., Li, M., Sun, Z., Sun, L., Huo, Q.: Robust table structure recognition with dynamic queries enhanced detection transformer. Pattern Recognition 144, 109817 (2023). https://doi.org/https://doi.org/10.1016/j.patcog.2023.109817, https://www.sciencedirect.com/science/article/pii/S0031320323005150
30. Xiao, B., Simsek, M., Kantarci, B., Alkheir, A.: Table structure recognition with conditional attention (03 2022)
31. Xiao, B., Simsek, M., Kantarci, B., Alkheir, A.A.: Rethinking detection based table structure recognition for visually rich documents (2023)