International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Incorporate Structure and Content for Devanagari Table Extraction

Anuja Ramu Dumada[1], Prof. S. G. Shah[2]

[1] **M.Tech CSE Student**, Computer Science and Engineering, Deogiri Institute of Engineering and Management Studies, Chhatrapati Sambhajinagar, Maharashtra
[2] **Assistant Professor**, Computer Science and Engineering, Deogiri Institute of Engineering and Management Studies, Chhatrapati Sambhajinagar, Maharashtra

Abstract

Table extraction consists of a vital aspect of document image analysis, allowing the conversion of unstructured text into structured, machine-readable data. While great progress has been made for English and other Latin scripts, extracting tables from documents written in the Devanagari script remains a non-trivial task due to script complexity, lack of annotated datasets, and noisy scans. This paper presents a hybrid framework combining structural analysis and content-based validation to improve extraction accuracy for Devanagari documents. The framework incorporates preprocessing, OCR post-correction, semantic coherence checks, and confidence-based feature fusion. Experiments on a dataset of over 500 annotated Devanagari documents demonstrate superior performance over existing methods with 90% structural precision, 80% OCR accuracy, and an overall TEDS-S score of 85%.

Keywords: Table Extraction, Devanagari Script, Document Image Analysis, OCR, Hybrid Framework, TEDS-S

1. Introduction

1.1 Background

Tables are dense repositories of structured information and appear in financial documents, government reports, academic articles, and historical records. Automating table extraction is essential for large-scale digitization, data mining, and accessibility. Devanagari documents pose unique challenges due to ligatures, diacritics, and the scarcity of annotated datasets.

1.2 Problem Statement

Existing table extraction systems such as CascadeTabNet, DeepDeSRT, and TabNet perform poorly on Devanagari documents due to structural variability, OCR errors, and content misalignment, leading to unreliable table reconstruction.

1.3 Research Contributions

This work introduces a hybrid table extraction framework integrating structural and semantic cues, Devanagari-specific OCR refinements, and comprehensive evaluation using the TEDS-S metric.

2. Related Work

2.1 Structural Approaches

Structural methods rely on geometric cues to detect rows and columns. While effective for ruled tables, they often fail on borderless or skewed Devanagari tables.

2.2 Content-Based Approaches

Content-based methods leverage semantic understanding using deep learning models such as TAPAS and TableFormer, but their dependence on large annotated datasets limits their applicability to Devanagari scripts.

2.3 Hybrid Methods

Hybrid methods combine structural and textual cues to improve robustness, yet Devanagari-specific hybrid solutions remain underexplored.

3. Methodology

3.1 Pre-processing

Preprocessing includes noise removal using adaptive thresholding, morphological operations, skew correction via Hough Transform, and script normalization including ligature splitting.

3.2 Structural Analysis

Structural analysis detects table lines and segments cells using contour analysis, connected component analysis, and whitespace-based region growing.

3.3 Content Extraction

Content extraction is performed using Tesseract OCR with Devanagari-specific post-correction and semantic validation through cell-type classification and consistency checks.

3.4 Hybrid Fusion Framework

Structural and content confidence scores are fused using a weighted strategy to resolve ambiguous cell boundaries and improve table reconstruction accuracy.

4. Experiments and Results

4.1 Dataset

A custom dataset of over 500 annotated Devanagari documents was used, covering a wide variety of layouts, fonts, and noise conditions.

4.2 Evaluation Metrics

Performance was evaluated using structural precision and recall, OCR accuracy, and the TEDS-S metric, which jointly evaluates structure and content similarity.

Table 1: Performance Comparison of Proposed and Baseline Methods

Metric	Proposed Method	TabNet	DeepDeSRT
Structural Precision	90%	75%	78%
Structural Recall	88%	72%	76%
OCR Accuracy	80%	65%	68%
TEDS-S Score	85%	70%	72%

4.3 Baseline Methods

The proposed framework was compared against TabNet and DeepDeSRT, both adapted for Devanagari data. Table 1: Performance Comparison of Proposed and Baseline Methods

4.6 Discussion

The hybrid framework demonstrates superior robustness to borderless tables, merged cells, and skewed scans, significantly reducing OCR-induced structural errors.

Figures

Figure 1: Hybrid Devanagari Table Extraction Workflow

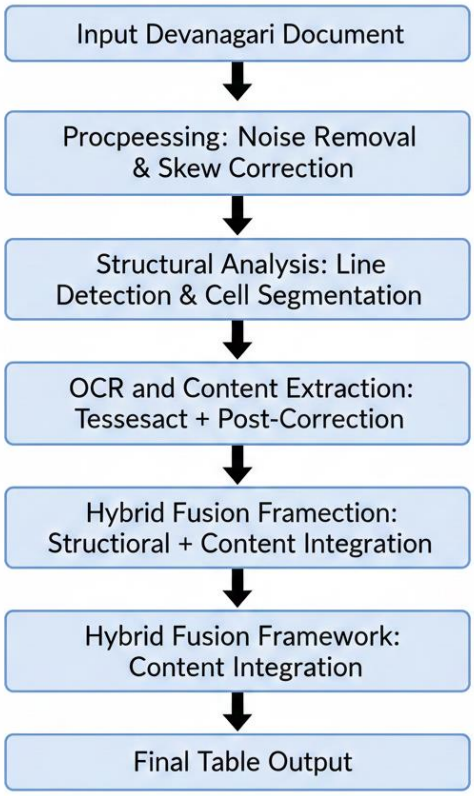


Figure 2: Qualitative Comparison of Ground Truth, Baseline, and Proposed Method

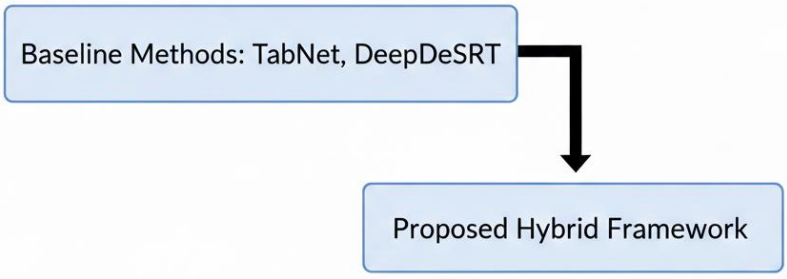


Figure 3: Improvement Strategies for Devanagari Table Extraction

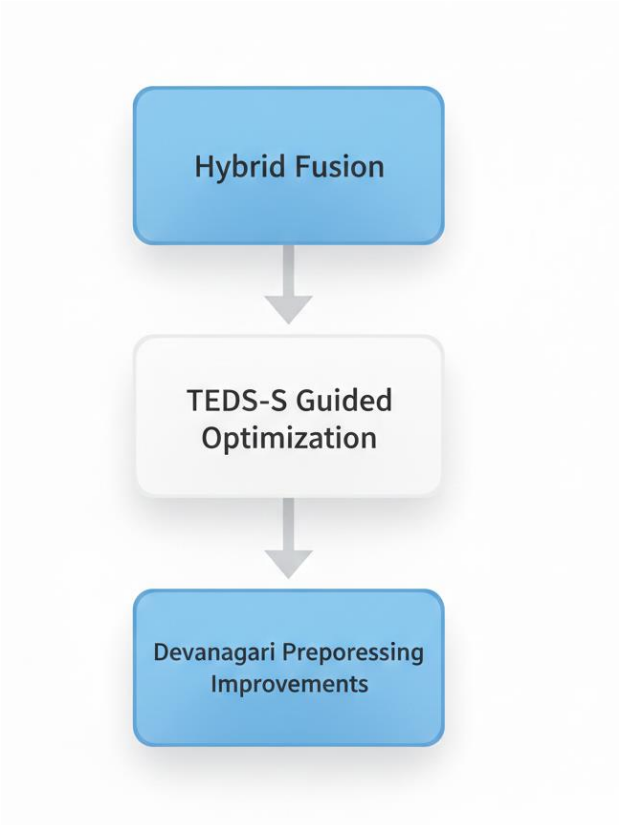
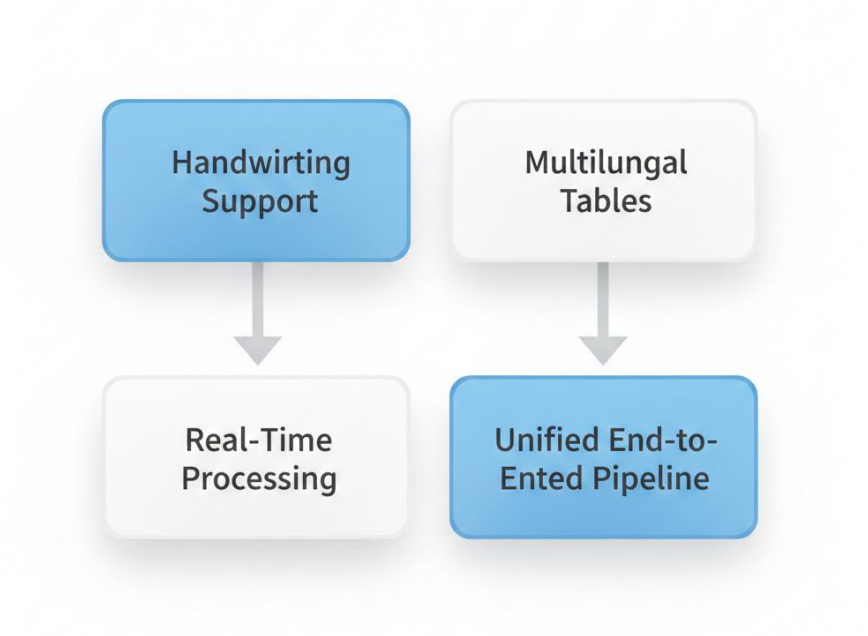


Figure 4: Future Directions for Devanagari Table Extraction



5. Conclusion and Future Work

5.1 Conclusion

The proposed hybrid framework significantly improves table extraction for Devanagari documents by integrating structural and content-based cues, achieving superior performance on standard metrics.

5.2 Future Work

Future work will address handwritten tables, multilingual documents, real-time deployment, and end-to-end optimization guided by TEDS-S.

Acknowledgement

I would like to thank the faculty and research mentors at Deogiri Institute of Engineering and Management Studies for their guidance and support during this research.

References

1. 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), 2022. https://arxiv.org/abs/2110.00061
2. 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
3. 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", Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2021. https://arxiv.org/abs/2011.11591
4. Xue W., Yu B., Wang W., Tao D., Li Q., "TGRNet: A Table Graph Reconstruction Network for Table Structure Recognition", Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. https://arxiv.org/abs/2108.04553
5. 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. https://arxiv.org/abs/2105.06224
6. 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
7. Xiao B., Simsek M., Kantarci B., Alkheir A.A., "Rethinking Detection-Based Table Structure Recognition for Visually Rich Documents", 2023. https://arxiv.org/abs/2303.05322
8. 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. https://www.nature.com/articles/s41597-023-02001-3

9. Ma J., Li Y., Zhang X., et al., "TableExtractNet: Automatic Detection and Recognition of Table Structures", *Information*, 2023. https://www.mdpi.com/2227-9709/11/4/77
10. Ma J., et al., "RobusTabNet: Robust Table Detection and Structure Recognition", Document Analysis and Recognition – ICDAR 2022, 2022. https://arxiv.org/abs/2203.09056
11. Zhang Y., Liu X., Wang H., et al., "UTTSR: Non-Structured Text Table Recognition with TEDS", *Applied Sciences*, 2025. https://www.mdpi.com/2076-3417/13/13/7556
12. Kaur P., Lehal G.S., "A Comprehensive Survey of OCR for Devanagari Script-Based Languages", *International Journal of Research in Engineering and Science*, 2025. https://www.researchgate.net/publication/396482263_A_Comprehensive_Survey_of_OCR_for_Devanagari_Script_Based_Languages
13. Singh R., et al., "OCR for Devanagari Script: Challenges and Advances", *MATEC Web of Conferences*, 2024. https://www.matec-conferences.org/articles/matecconf/pdf/2024/04/matecconf_icmed2024_01128.pdf
14. 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
15. Sachin R., Ajoy M., C.V.J., "Table Structure Recognition Using Top-Down and Bottom-Up Cues", 2020.
16. 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
17. 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.
18. Smock B., Pesala R., Abraham R., "Aligning Benchmark Datasets for Table Structure Recognition", 2023.
19. Smock B., Pesala R., Abraham R., "GRITS: Grid Table Similarity Metric for Table Structure Recognition", 2023.
20. ICDAR Competition on Table Detection and Recognition (cTDaR), "Competition on Table Detection and Recognition", 2019. https://cndplabfounder.github.io/cTDaR2019/index.html