Library Progress International
Vol.44 No. 3, Jul-Dec 2024: P.1919-1931

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

Performance Comparison of Tesseract and Google Document AI in Punjabi Newspapers Digitization

[1]Atul Kumar*, [2]Gurpreet Singh Lehal

[1]Department of Computer Science, R.G.M. Govt. College Joginder Nagar, Mandi, 175015, India
E-mail: atulkmr02@gmail.com
ORCID id: https://orcid.org/0000-0002-7665-1892
[2]Department of Computer Science, Punjabi University, Patiala, 147002, India
E-mail: gslehal@gmail.com
ORCID id: https://orcid.org/0000-0001-6152-8050

How to cite this article: Atul Kumar, Gurpreet Singh Lehal (2024) Performance Comparison of Tesseract and Google Document AI in Punjabi Newspapers Digitization. *Library Progress International,* 44(3), 1919-1931

ABSTRACT

This paper focuses on the digitization of Punjabi newspapers through Optical Character Recognition (OCR) technology, a comparative analysis was conducted between Google Document AI and Tesseract OCR solutions. Punjabi newspapers, with their complex layouts, non-standard fonts, and linguistic nuances, pose challenges for OCR systems. The study aimed to evaluate the out-of-the-box performance of OCR solutions in accurately extracting text from Punjabi newspaper scans. Utilizing a benchmarking experiment with a dataset comprising Punjabi newspaper segments, the research addressed questions regarding the comparative performance of Google Document AI and Tesseract in handling Punjabi text. The methodology involved image enhancement, layout analysis, and OCR execution, with qualitative and quantitative analyses conducted to assess precision and reliability. While Tesseract demonstrated competitive performance, Google Document AI exhibited superior accuracy, highlighting the potential of server-based OCR solutions for handling diverse document types. The mask RCNN model is used to extract the layout of newspapers using a layout parser. The findings reveal that while Tesseract demonstrates competitive performance, Google Document AI exhibits superior accuracy. We have performed the text extraction on newspaper segments that are extracted from newspaper images. Specifically, Tesseract achieved an accuracy of 97.20% at the word level and 92.48% at the character level, whereas Google API performed better with an accuracy of 98.86% at the word level and 95.62% at the character level. These findings contribute to the advancement of OCR technology in the context of Punjabi newspaper digitization, facilitating broader access to historical Punjabi texts for scholarly research.

KEYWORDS

OCR, Newspaper, Tesseract, Google API, Digital.

INTRODUCTION

Optical Character Recognition (OCR) technology has emerged as a critical tool in the digitization of Punjabi newspapers, offering researchers unprecedented access to historical texts for social scientific and humanities research. The ability to automatically extract text from digital images holds immense promise for uncovering insights into Punjabi culture, language, and societal dynamics embedded within these publications. However, the effectiveness of OCR solutions, particularly in handling the unique characteristics of Punjabi newspapers, remains a subject of investigation.

In this research paper, we aim to conduct a comparative analysis of two prominent OCR solutions: Google Document AI and Tesseract. While OCR technology has undergone significant advancements in recent years, the performance of OCR systems on Punjabi newspaper scans, which often feature complex layouts, non-standard fonts, and noise, warrants further examination. Specifically, we focus on evaluating the out-of-thebox performance of Google Document AI and Tesseract in accurately extracting text from Punjabi newspaper images.

Historically, general OCR processors like Tesseract have faced challenges in achieving high accuracy rates when confronted with real-world complexities such as shading, blur, and non-standard fonts commonly found in Punjabi newspapers. Moreover, the linguistic nuances of Punjabi pose additional obstacles for OCR systems, especially those trained primarily in Western languages. The recent emergence of server-based OCR solutions, such as Google Document AI, has raised expectations for improved performance in handling diverse document types, including those in non Western languages.

This study seeks to address the following research gaps and questions:

- How does the performance of Google Document AI compare with Tesseract in accurately extracting text from Punjabi newspaper scans?
- Punjabi newspapers have sophisticated layouts. Currently, there is no such mechanism that addresses this issue. We addressed this issue in our research.
- We address the issue of newspaper image enhancements to improve the recognition.
- How do Google Document AI and Tesseract perform in handling Punjabi text and what implications does this have for scholars working with newspaper recognition systems?

To achieve these objectives, we conducted a benchmarking experiment comparing the performance of Google Document AI and Tesseract on a dataset comprising Punjabi newspaper scans. The experiment aimed to provide statistically meaningful measurements of OCR accuracy, enabling researchers to make informed decisions regarding the selection of OCR solutions for their digitization projects.

1. Literature Survey

The history of OCR dates back to the mid-20th century, with early efforts focused on mechanical devices for reading characters. The advent of computers in the latter half of the century paved the way for the development of electronic OCR systems. Tesseract is a widely used open-source Optical Character Recognition (OCR) engine that was originally developed by Hewlett Packard and later supported by Google (Smith, 2007). (Edupuganti et al., 2021) proposed the development of a mobile application utilizing the Google Vision library to empower visually impaired individuals by enabling text recognition, detection, and conversion to speech, thereby facilitating independent medicine identification and consumption. (Drobac et al., 2019) conducted experiments on Optical Character Recognition (OCR) of historical Finnish newspapers and journals demonstrate promising results, with the mixed model achieving a 95% Character Accuracy Rate on the Finnish test set, surpassing previous results on this dataset. (Zhu et al., 2022) released a dataset of 3000 annotated newspaper images from 21 U.S. states, proposed layout segmentation as a preprocessing step for OCR, and established a thorough evaluation protocol for layout segmentation and end-to-end OCR. (Almutairi & Almashan, 2019) proposed a deep learning system for semantic segmentation of the key newspaper elements and used the instance segmentation method mask R-CNN to build a language-independent model that logically deconstructed a newspaper page's raw image into its main elements based solely on its visual features. (Gupta et al., 2007) used error diffusion binarization for binarization, pre-filtering, and post-binarization denoising. Methods were compared using ABBYY FineReader 7.1 SDK and performed best on 12 pages from six newspapers of diverse quality. (Robby et al., 2019) addressed the challenge of Optical Character Recognition (OCR) for non-Latin scripts, focusing on Javanese characters. A dataset comprising 5880 characters was collected and trained using various methods with Tesseract OCR tools. The implemented models, optimized with boundary box configurations, achieved a peak accuracy of 97.50%, demonstrating promise for mobile application integration. (Gemelli et al., 2024) conducted a thorough investigation and comparison of the most frequently utilized datasets for layout analysis, with a specific emphasis on those pertaining to scientific publications and gave an executive summary of the most popular approaches developed for and evaluated with these datasets. (Martínek et al., 2020) comprised page layout analysis, which encompassed text block and line segmentation, as well as OCR. Segmentation employed fully convolutional networks, while OCR used recurrent neural networks, both acknowledged as cuttingedge. Experiments were carried out to determine the most effective methods for achieving high performance with limited training data. (Koistinen et al., 2020) presented efforts to enhance the optical character recognition (OCR) quality of historical Finnish newspapers at the National Library of Finland. Using a 500,000-word sample, they compare OCR results between ABBYY FineReader and Tesseract, achieving significant improvements in precision, recall, and character accuracy rates with Tesseract. (Kaur et al., 2019) introduced a system for Gurumukhi script newspaper recognition, employing four feature extraction techniques and four classification methods. Utilizing data from

major newspapers, the system achieved recognition accuracy of 96.19% with a combination of zoning, diagonal, and parabola curve fitting features, and 95.21% with a partitioning strategy of 70% training and 30% testing data. (Bansal et al., 2014) segmented page images from English newspaper, labeling various blocks such as headlines, subheadings, captions, images, and text. They used a fixed-point model with SVM and KLR as prediction functions, achieving 95% accuracy in block labeling. (Ghosh, 2023) introduced the first Bengali script newspaper text recognition technology. The newspaper piece has image and text sections, text lines, words, and characters. Different methods were used to detect character qualities. Newspaper characters were distinguished by an SVM classifier using feature vectors. The suggested system had 97.78% text recognition accuracy on a self generated dataset. (Kohli et al., 2022) presented model, named the J&M model, focuses on detecting text from handwritten images. Implemented in Python using the MNIST database of handwritten digits, the research attains impressive results, with a training accuracy of 99.5% and testing accuracy of 99%, alongside a training loss of 1.5%. (Ye & Doermann, 2015) analyzed text detection and recognition in color imagery, categorizing techniques, and addressing subproblems such as localization and segmentation. It also examines challenges like degraded text enhancement and processing multi-oriented, perspectively distorted, and multilingual text, offering insights into benchmark datasets and comparing the performance of leading approaches.In case of layout analysis, (Singh & Kumar, 2014) proposed approach combines bottom-up region growing and top-down segmentation methods for document layout analysis, effectively leveraging both approaches simultaneously.

2. Methodology

The existing technique for recognizing text from Punjabi newspapers is to run the OCR directly. Because newspapers have complex layouts, these techniques produce erroneous findings. There is currently no recognition system for Punjabi newspapers that takes into account the aforementioned constraint.

Our methodology includes newspaper layout analysis, segmentation, and image enhancement of newspaper segments, followed by a comprehensive evaluation of Tesseract OCR and Google Docs OCR using a standardized test dataset comprising 500 segments extracted from Punjabi newspaper images during segmentation. Figure 1 illustrates the adopted methodology for text extraction. Both qualitative and quantitative analyses are conducted to assess the precision and reliability of the OCR systems. Qualitative evaluation involves visual inspection of OCR outputs, while quantitative evaluation includes metrics such as character accuracy, word accuracy, and processing speed.

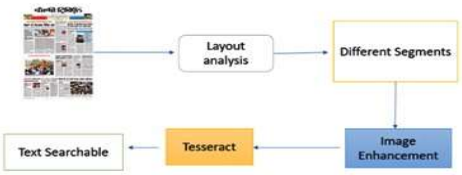

Fig. 1. Methodology for Text Extraction from Punjabi Newspapers

3.1 Layout Analysis

To extract the segments, the approach begins with passing the newspaper image via the layout analysis module. We utilized a layout parser to identify the layouts of newspapers. Layout analysis plays a crucial role in understanding the structure and content arrangement of newspapers. We present a detailed analysis of the layout of a newspaper using the mask_rcnn_R_50_FPN_3x model configuration. The backbone of the model consists of a ResNet-50 with a Feature Pyramid Network (FPN) architecture as shown in Figure 2. This backbone is adept at capturing hierarchical features at different scales, which is crucial for analyzing the diverse elements present in a newspaper layout. The RPN (Region Proposal Network) module is responsible for generating region proposals for potential objects in the newspaper. It utilizes a standard RPN head and processes input features from different pyramid levels (p2 to p6) to propose regions of interest. The model's ROI heads are designed to refine and classify the proposed regions. With standard ROI heads, the model identifies objects and segments instances within the proposed regions. The ROI box head performs bounding box regression, while the ROI mask head generates pixel-wise masks for each detected layout of the newspaper along with the objectiveness score.

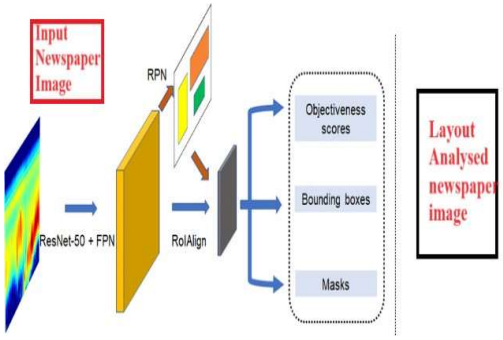

Fig. 2. Architecture of Mask RCNN

3.1.1. Configuration of Model

The config_path parameter specifies the path to the model configuration file in the Layout Parser model catalog, loading the mask_rcnn_R_50_FPN_3x model configuration. Mask RCNN R50 architecture using Detectron 2 which is a popular deep learning framework developed by Facebook AI Research for computer vision tasks, including object detection, instance segmentation, and keypoint detection. The label_map parameter maps class labels used by the model to the names of layout regions they represent, with 1 representing text regions and 2 representing image regions. Additionally, the extra_config parameter allows for additional configuration options for the deep learning model, setting the minimum score threshold for object detection to 0.1, ensuring that only objects with a score above this threshold will be considered during layout analysis. On applying the model on various Punjabi newspaper images, we got the result as shown in Figure 3.

Fig. 3. Output image of Punjabi Newspaper after Layout analysis

3.2. Image Enhancement

After the layout analysis, various segments are extracted from the newspaper. Following this, various portions are extracted and improved to boost their quality before being sent to Tesseract. We enhanced images by applying dilation, difference, and normalization operations to portions of newspapers. The workflow of Newspaper image enhancement is shown in Figure 4.

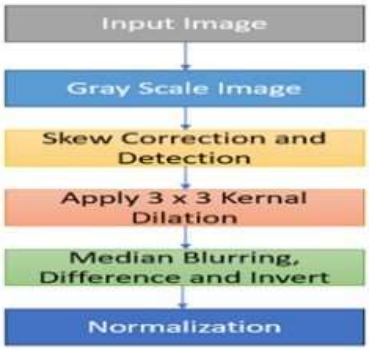

Fig. 4. Workflow for enhancement of Newspaper Image

The input newspaper image is transformed into a grayscale image. A skew is discovered on a newspaper image

and fixed. Then we apply 3 X 3 dilation on the kernel. Following median blurring, difference, and inversion operations are conducted on the image, followed by normalization to get the final enhanced image. The enhanced image is shown in Figure 5.

Fig. 5. a) Noisy Image **b)** Results of Image Enhancements on Punjabi Newspaper Segments

3.3. Text Recognition using Tesseract and Google OCR

We have chosen two OCR services Tesseract and Google API for benchmarking experiments with the newspaper image segments of the Punjabi language. Table 1 shows the features of Tesseract OCR and Google API. Table 2 contains information about the quality of scanned newspaper images. It includes the following information:

1. **Resolution:** The majority of newspaper images (300) have high or medium resolution.
2. **Brightness:** A significant portion of the documents (23) have low brightness.
3. **Common Issues:** Blurred images (15), improper scans resulting in black tones (10), and skewed scans (16) are the most frequent issues encountered.
4. **Old Newspapers:** A specific category of parser (Shen et al., 2021) library designed for identifying the documents (26) is identified as "Old Newspaper" layout for these newspaper images. The library is with low resolution and noise built on Faster RCNN and Mask RNN models, trained on the extensive PRIMA Layout and PublayNet datasets (Zhong et al., 2019).
5.

Table 1. Features of Tesseract OCR and Google API

S.No.	Feature	Tesseract	Google Cloud Vision API
1	Type	Open-source	Cloud-based (Google Cloud)
2	Cost	Free	Free tier with limited usage, then pay-as-you-go
3	Offline capability	Yes	No (requires internet connection)
4	Language support	Over 100 languages	Similar to Tesseract
5	Basic text extraction	Good	Good
6	Layout analysis	Limited	Advanced (tables, paragraphs, logos)
7	Handwriting recognition	Limited	Better accuracy
8	Object detection	No	Yes
9	Scalability	Limited	High scalability

Category	Feature	Count
Resolution	High	200
	Medium	100
	Low	100
Brightness	Low	23
Blurred	Low Resolution	15
Improper Scan	Blackish tone	10
Skewed	Due to scanning	16
Old Newspaper	Low resolution and noisy	26
	Total	500

10	Best for	Basic OCR needs offine processing	Advanced features, complex images, large volumes

Table 2. Characteristics of Sample of Newspaper Segments

3.3.1 Algorithm for Text Extraction

1. Take a Punjabi newspaper image.
2. Perform Layout analysis using a layout parser.
3. Perform Skew correction and Dilation over 3X3 kernel on the segment extracted.
4. Apply a median filter on the dilated image and find the difference between the blurred and original image
5. Invert the image and normalize the segment image.
6. Employ 'image_to_string()' function from the Tesseract module and Google Document API to execute OCR on the image.
7. Specify the language model for recognizing Punjabi text by setting the 'lang' parameter to 'Pan'.
8. Optimize the OCR engine's settings by incorporating additional configurations through the 'config' parameter.
9. Fetch the text recognized by the OCR (Tesseract and Google) engine.

The proposed technique solves the issues of identifying text from Punjabi newspapers by using a holistic approach that includes layout analysis, segmentation, image improvement, and OCR benchmarking. The approach ensures precise extraction of newspaper segments by applying advanced models for layout analysis, such as mask_rcnn_R_50_FPN_3x. Image enhancement techniques like dilation and median blurring improve text

legibility while eliminating errors caused by noise or low-quality scans. Benchmarking tests with Tesseract and Google API are used to undertake both qualitative and quantitative analyses of OCR performance, including character and word accuracy. This comprehensive strategy attempts to improve the accuracy of text recognition in Punjabi newspapers, resulting in more dependable results by efficiently resolving layout and image quality difficulties.

3. Experiments, Results and Discussions

The experiments used Tesseract version 4.1.1.2 and Python version 3.11. The GPU used is the V100, and the RAM is 16GB. The experiments involve Punjabi newspapers collected from various sources such as websites and libraries, totaling around 300 newspaper images. Tesseract OCR and Google OCR were chosen due to their support for the Punjabi language. After the segmentation of these newspapers, approximately 100 segments of newspaper clips were selected for further processing and evaluation. The experiment was designed to evaluate out-of-the-box performance, where newspaper segments were processed using the two OCR engines. Figure 6 shows the visual representation of the recognition of Punjabi newspaper images.

Accuracy was measured by developing a Python tool to compare the actual words in the image text file with the recognized text file. This tool calculates the word error rate and character error rate. The chosen model, Mask RCNN (Mask Regionbased Convolutional Neural Network), for layout analysis in the text extraction methodology from Punjabi newspapers, is rational due to its robust capabilities in instance segmentation, hierarchical feature extraction, and precise region proposal and classification. With its ResNet-50 backbone and Feature Pyramid Network (FPN) architecture, Mask RCNN effectively captures hierarchical features at different scales, essential for analyzing the diverse elements present in a newspaper layout. Its Region Proposal Network (RPN) generates accurate region proposals, while the ROI heads refine and classify these proposed regions, aiding in accurate segmentation and layout understanding. Moreover, Mask RCNN's ability to generate pixel-wise masks for each detected layout element ensures precise boundaries, facilitating subsequent processing steps such as image enhancement and text recognition. The results of the comparative performance analysis reveal notable differences between Tesseract OCR and Google Docs OCR in terms of accuracy and reliability. In the results and discussion section, we focus on the testing of the Cloud Vision API and Tesseract-based systems using a significant number of Punjabi newspaper segments to ensure variance in the data.

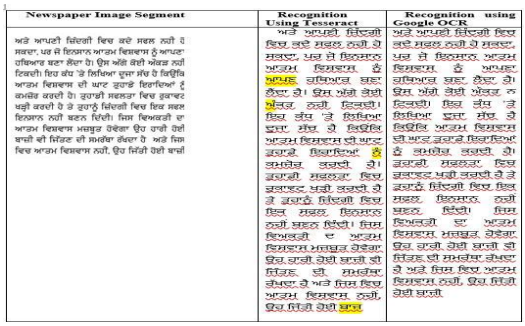

Fig 6. Text extraction using Tesseract and Google OCR.

Figure 6 shows yellow highlights on words, which demonstrate OCR misrecognition. Tesseract OCR incorrectly recognizes three words in this example, whereas Google OCR provides 100% accuracy on this image. Table 3 presents the comparison of the performance of these systems using 100 segments. Each set includes information on the number of segments, the total number of words and characters in those segments, the number of words and characters recognized by the OCR system, and the resulting accuracy percentages for both words and characters. Tesseract provides an accuracy of 97.20% at the word level and 92.48% at the character level. Google API performs better giving an accuracy of 98.86% at the word level and 95.62% at the character level. Figure 7 shows the error rate on 100 segments of Punjabi Newspapers.

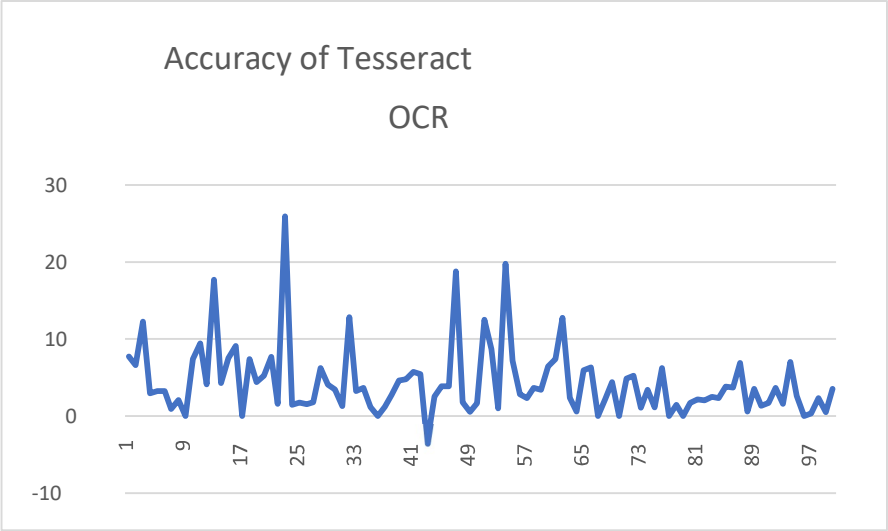

Fig 7. Error rate of Tesseract OCR on 100 segment images

Number of segments	Number of words	Number of Characters	Recognized words	Recognized Characters	Accuracy(word)	Accuracy (Characters)
100	12604	28654	12252	26500	97.20%	92.48%
100	12604	28654	12461	27400	98.86%	95.62%

Tesseract OCR demonstrates competitive performance in extracting text from newspaper images, achieving high character and word accuracy rates across the test dataset. However, it is observed that Tesseract OCR may exhibit variations in performance depending on factors such as image quality, text layout, and language complexity. On the other hand, Google Docs OCR exhibits consistent performance but may require access to cloud services and incur additional costs. Table 4 displays the outcomes of text recognition on various images employing Tesseract and Google OCR techniques on Punjabi newspaper segments. Along with error rate, The results indicate that Google API tends to demonstrate superior accuracy compared to Tesseract, with a majority of images yielding a 0% error rate for Google API, whereas Tesseract exhibits higher error rates. A selection of sample segments analyzed using Tesseract and Google OCR is presented in Table 4 for reference and analysis. It has been found from the study that some similarly shaped characters are misrecognized by the Tesseract OCR than Google API. Based on the findings, we found that the error rates of Tesseract OCR on Punjabi newspaper segments vary, with some segments showing significantly higher error rates compared to Google API. Image quality, text layout, and language complexity significantly affect OCR performance. Tesseract OCR exhibit variations in performance based on these factors, leading to higher error rates in some cases. On the other hand, Google Docs OCR demonstrates consistent performance across different image qualities and text layouts. It provides advanced features such as layout analysis, handwriting recognition, and object detection, making it suitable for processing complex images and large volumes of text. However, Tesseract OCR, being open-source and offline-capable, is suitable for basic OCR needs and offline processing. It provides good performance in extracting text from newspaper images, achieving competitive accuracy rates across the test dataset. Further research could explore the development of hybrid OCR systems that combine the strengths of Tesseract and Google Docs OCR to achieve even higher accuracy rates. Investigating the impact of different preprocessing techniques on OCR performance could also yield valuable insights for improving accuracy in various scenarios. Overall, the rationality of the chosen model lies in its ability to effectively address the complexities of Punjabi newspaper layouts, enhance image quality, and provide a comprehensive evaluation of OCR performance, ultimately leading to improved accuracy in text recognition

Table 3. Comparison between Tesseract OCR and Google OCR
Table 4. Word error rates by Tesseract and Google API for Punjabi Newspaper Image segments

Image No.	Total Words	Tesseract OCR	Google API	Error rate Tesseract	Error Rate Google OCR
01	142	131	142	7.746479	0
02	151	141	151	6.622517	0
03	57	50	56	12.2807	1.75438596
04	136	132	136	2.941176	0
05	186	180	186	3.225806	0
06	124	120	122	3.225806	1.61290323
07	108	107	107	0.925926	0.92592593
08	145	142	145	2.068966	0
09	131	131	131	0	0
10	108	100	108	7.407407	0
11	106	96	102	9.433962	3.77358491
12	73	70	73	4.109589	0
13	147	121	142	17.68707	3.40136054
14	186	178	183	4.301075	1.61290323
15	120	111	118	7.5	1.66666667
16	99	90	97	9.090909	2.02020202
17	141	141	141	0	0
18	54	50	53	7.407407	1.85185185
19	136	130	136	4.411765	0
20	153	145	153	5.228758	0
21	65	60	65	7.692308	0
22	63	62	63	1.587302	0
23	54	40	54	25.92593	0
24	140	138	140	1.428571	0
25	173	170	171	1.734104	1.15606936
26	132	130	132	1.515152	0
27	112	110	110	1.785714	1.78571429
28	96	90	95	6.25	1.04166667
29	171	164	169	4.093567	1.16959064
30	117	113	115	3.418803	1.70940171
31	155	153	152	1.290323	1.93548387
32	70	61	70	12.85714	0
33	154	149	150	3.246753	2.5974026
34	110	106	106	3.636364	3.63636364
35	172	170	171	1.162791	0.58139535
36	78	78	75	0	3.84615385
37	161	159	161	1.242236	0
38	181	176	173	2.762431	4.4198895
39	173	165	171	4.624277	1.15606936
40	126	120	126	4.761905	0
41	122	115	122	5.737705	0

Image No.	Total Words	Tesseract OCR	Google API	Error rate Tesseract	Error Rate Google OCR
42	110	104	107	5.454545	2.72727273
43	112	116	110	-3.571429	1.78571429
44	118	115	115	2.542373	2.54237288
45	180	173	178	3.888889	1.11111111
46	52	50	51	3.846154	1.92307692
47	80	65	80	18.75	0
48	164	161	162	1.829268	1.2195122
49	185	184	182	0.540541	1.62162162
50	61	60	58	1.639344	4.91803279
51	80	70	78	12.5	2.5
52	69	63	65	8.695652	5.79710145
53	98	97	95	1.020408	3.06122449
54	81	65	80	19.75309	1.2345679
55	167	155	167	7.185629	0
56	177	172	172	2.824859	2.82485876
57	172	168	172	2.325581	0
58	136	131	135	3.676471	0.73529412
59	178	172	178	3.370787	0
60	139	130	139	6.47482	0
61	54	50	54	7.407407	0
62	141	123	139	12.76596	1.41843972
63	125	122	122	2.4	2.4
64	163	162	163	0.613497	0
65	152	143	152	5.921053	0
66	143	134	143	6.293706	0
67	121	121	120	0	0.82644628
68	178	174	178	2.247191	0
69	136	130	132	4.411765	2.94117647
70	188	188	188	0	0
71	184	175	182	4.891304	1.08695652
72	95	90	95	5.263158	0
73	95	94	93	1.052632	2.10526316
74	148	143	148	3.378378	0
75	267	264	266	1.123596	0.37453184
76	80	75	79	6.25	1.25
77	118	118	118	0	0
78	142	140	140	1.408451	1.4084507
79	65	65	63	0	3.07692308
80	117	115	115	1.709402	1.70940171
81	91	89	85	2.197802	6.59340659
82	98	96	95	2.040816	3.06122449
83	80	78	78	2.5	2.5
84	129	126	129	2.325581	0
85	105	101	101	3.809524	3.80952381

Image No.	Total Words	Tesseract OCR	Google API	Error rate Tesseract	Error Rate Google OCR
86	81	78	80	3.703704	1.2345679
87	58	54	58	6.896552	0
88	169	168	168	0.591716	0.59171598
89	85	82	85	3.529412	0
90	151	149	145	1.324503	3.97350993
91	174	171	172	1.724138	1.14942529
92	137	132	137	3.649635	0
93	188	185	188	1.595745	0
94	57	53	56	7.017544	1.75438596
95	115	112	113	2.608696	1.73913043
96	25	25	25	0	0
97	311	310	311	0.321543	0
98	85	83	82	2.352941	3.52941176
99	210	209	210	0.47619	0
100	56	54	56	3.571429	0

4. Conclusion

In conclusion, we presented a methodology for the recognition of Punjabi newspapers. We have used the Mask RCNN model for detecting the layout of Punjabi newspapers. We have presented techniques for image enhancements of Punjabi newspaper segments. The study paper also provides a thorough comparison investigation of Tesseract OCR and Google API in extracting text from Punjabi newspaper photos, focusing on their precision and dependability. Although OCR algorithms have historically struggled with non-standard typefaces and linguistic complexities, Tesseract OCR shows strong performance by reaching high accuracy rates across the sample. Google API demonstrates high accuracy in certain situations; however, it could result in extra expenses and dependence on cloud services. The results have important significance for researchers working on digitizing Punjabi newspapers. This offer useful insights into the effectiveness of OCR technology and can help guide future research efforts to improve OCR performance for complicated scripts such as Punjabi.

There are also some shortcomings in the research process of this paper, which are worthy of further study. For example, we focused solely on comparing the performance of two OCR solutions, Tesseract and Google API. Including additional OCR solutions in the comparison could provide a more comprehensive understanding of OCR performance. The evaluation focused specifically on the recognition of Punjabi text. While this is important for the context of the study, the performance of OCR solutions may vary for other languages and scripts. Generalizing the findings beyond Punjabi may require further investigation.

Future research might look into the performance of other OCR methods, broaden the dataset to include more diverse newspaper layouts and content and use more subtle evaluation measures to capture finer nuances of OCR accuracy. Furthermore, initiatives to automate image enhancement procedures and address cost concerns connected with cloud-based OCR solutions may improve the feasibility and accessibility of OCR technology for digitizing historical writings in a variety of languages

References

Almutairi, A., & Almashan, M. (2019). Instance Segmentation of Newspaper Elements Using Mask R-CNN. In *18th IEEE International Conference On Machine Learning And Applications*. https://doi.org/10.1109/icmla.2019.00223

Bansal, A., Chaudhury, S., Roy, S. D., & Srivastava, J. (2014). Newspaper Article Extraction Using Hierarchical Fixed Point Model. Document Analysis Systems(DAS), 11th IAPR International Workshop. IEEE. https://doi.org/10.1109/das.2014.42

Drobac, S., Kauppinen, P., & Lindén, K. (2019). Improving OCR of historical newspapers and journals published in Finland. Proceedings of the 3nd International Conference on Digital Access to Textual Cultural Heritage. https://doi.org/10.1145/3322905.3322914

Edupuganti, S. A., Koganti, V. D., Lakshmi, C. S., Kumar, R. N., & Paruchuri, R. (2021). Text and Speech Recognition for Visually Impaired People using Google Vision. 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC). https://doi.org/10.1109/icosec51865.2021.9591829

Gemelli, A., Marinai, S., Pisaneschi, L., & Santoni, F. (2024). Datasets and annotations for layout analysis of scientific articles. *International Journal on Document Analysis and Recognition (IJDAR)*. https://doi.org/10.1007/s10032-024-00461-2

Ghosh, R. (2023). Newspaper text recognition in Bengali script using support vector machine. Multimedia Tools and Applications, 83(11), 32973–32991. https://doi.org/10.1007/s11042-023-16862-0

Gupta, M. R., Jacobson, N. P., & Garcia, E. K. (2007). OCR binarization and image pre-processing for searching historical documents. Pattern Recognition, 40(2), 389–397. https://doi.org/10.1016/j.patcog.2006.04.043

Kaur, R. P., Jindal, M. K., & Kumar, M. (2019). Recognition of newspaper printed in Gurumukhi script. Journal of Central South University, 26(9), 2495–2503. https://doi.org/10.1007/s11771-019-4189-1

Kohli, H., Agarwal, J., & Kumar, M. (2022). An improved method for text detection using Adam optimization algorithm. Global Transitions Proceedings, 3(1), 230–234. https://doi.org/10.1016/j.gltp.2022.03.028

Koistinen, M., Kettunen, K., & Kervinen, J. (2020). How to Improve Optical Character Recognition of Historical Finnish Newspapers Using Open Source Tesseract OCR Engine – Final Notes on Development and Evaluation. In Lecture notes in computer science (pp. 17–30). https://doi.org/10.1007/978-3-030-66527-2_2

Martínek, J., Lenc, L., & Král, P. (2020). Building an efficient OCR system for historical documents with little training data. *Neural Computing and Applications*, *32*(23), 17209–17227. https://doi.org/10.1007/s00521-020-04910-x

Robby, G. A., Tandra, A., Susanto, I., Harefa, J., & Chowanda, A. (2019). Implementation of Optical Character Recognition using Tesseract with the Javanese Script Target in Android Application. Procedia Computer Science, 157, 499–505. https://doi.org/10.1016/j.procs.2019.09.006

Shen, Z., Zhang, R., Dell, M., Lee, B. C. G., Carlson, J., & Li, W. (2021). LayoutParser: A Unified Toolkit for Deep Learning Based Document Image Analysis. In Lecture notes in computer science (pp. 131–146). https://doi.org/10.1007/978-3-030-86549-8_9

Singh, V., & Kumar, B. (2014). Document layout analysis for Indian newspapers using contour based symbiotic approach. 2014 International Conference on Computer Communication and Informatics.https://doi.org/10.1109/iccci.2014.6921723

Smith, R. (2007). An overview of the Tesseract OCR engine. Proceedings of the International Conference on Document Analysis and Recognition.https://doi.org/10.1109/icdar.2007.4376991

Ye, Q., & Doermann, D. (2015). Text Detection and Recognition in Imagery: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(7), 1480–1500. https://doi.org/10.1109/tpami.2014.2366765

Zhong, X., Tang, J., & Yepes, A. J. (2019). PubLayNet: Largest Dataset Ever for Document Layout Analysis. Proceedings of the International Conference on Document Analysis and Recognition, ICDAR,. https://doi.org/10.1109/icdar.2019.00166

Zhu, W., Sokhandan, N., Yang, G., Martin, S., & Sathyanarayana, S. (2022). DocBed: A Multi-Stage OCR Solution for Documents with Complex Layouts. Proceedings of the AAAI Conference on Artificial Intelligence, 36(11), 12643–12649. https://doi.org/10.1609/aaai.v36i11.21539