



A Hybrid Approach for Complex Layout Detection of Newspapers in Gurumukhi Script Using Deep Learning





Atul Kumar^{1*} and Gurpreet Singh Lehal²



¹Department of Computer Science, R.G.M. Govt. College Joginder Nagar, Mandi, Himachal Pradesh, India;

²Department of Computer Science, Punjabi University, Patiala, Punjab, India

E-mail/Orcid Id:

AK,  atulkmr02@gmail.com,  <https://orcid.org/0000-0002-7665-1892>; GSL,  gslehal@gmail.com,  <https://orcid.org/0000-0001-6152-8050>

Article History:

Received: 30th Aug., 2023
Accepted: 06th Nov., 2023
Published: 30th Nov., 2023

Keywords:

Deep Learning, CNN,
Gurumukhi, Layout,
Newspapers

How to cite this Article:

Atul Kumar and Gurpreet Singh Lehal (2023). A Hybrid Approach for Complex Layout Detection of Newspapers in Gurumukhi Script Using Deep Learning. *International Journal of Experimental Research and Review*, 35(Spl.), 34-42.
DOI : <https://doi.org/10.52756/ijerr.2023.v35spl.004>

Abstract: Layout analysis is the crucial stage in the recognition system of newspapers. A good layout analysis results in better recognition results. The complexity of newspaper layout structures poses a formidable challenge in digitization. The intricate arrangement of text, images, and various sections within a newspaper demands sophisticated algorithms and techniques for accurate layout detection. The paper introduces a diverse set of methodologies from existing literature, highlighting the evolution of techniques for newspaper layout analysis. In this paper, we present a novel method to detect the complex layout of newspapers in the Gurumukhi script by using a hybrid approach. The method developed consists of two parts. In the first part, we proposed an algorithm to remove pictures and graphics from Punjabi newspaper images that involve various image preprocessing tasks based on binarization, finding contours, and erosion on the image to remove the graphics from the image. This method removes pictures from complex non-Manhattan layouts. We have tested this algorithm on 100 newspaper images, giving an accuracy of 96.22%. In the second part, a dataset of 500 newspapers was created with images labeled with five classes on which the model was trained. Finally, we have trained the deep-learning model based on a convolutional network to detect the columns of text in newspapers. We have used four different architectures of CNN and compared their performance based on different metrics such as precision, recall, and F1 score. We have tested this method on a number of newspapers in the Gurumukhi script. We have achieved an accuracy of 95.53% with this approach.

Introduction

Newspaper digitization has gained significant importance in today's context, addressing the need to make newspapers more accessible and user-friendly. With newspapers being a vital information source, the challenge lies in efficiently accessing specific newspapers and the desired information within a vast collection of old newspapers. Therefore, the essential requirement for newspaper digitization arises, enabling users to retrieve information from their own workspace. The first step in the digitization process is to conduct a layout analysis of newspapers. As the newspaper is quite complex, detecting the correct layout is challenging. We have done

the layout analysis of a newspaper written in the Punjabi language. The primary factor driving this initiative is the predominant use of other Gurumukhi scripts for many reasons. Gurumukhi serves as the primary script for Punjab, where it is not only the first language but also ranks as the world's most spoken language. Hence, due to the widespread readership of newspapers in the Gurumukhi script, the conversion of old newspapers into digital formats becomes imperative. This transformation facilitates the accessibility of historical newspaper content on computers and other digital platforms.

Various methodologies have been proposed in the literature to detect the newspaper's layout. Chowdhury et



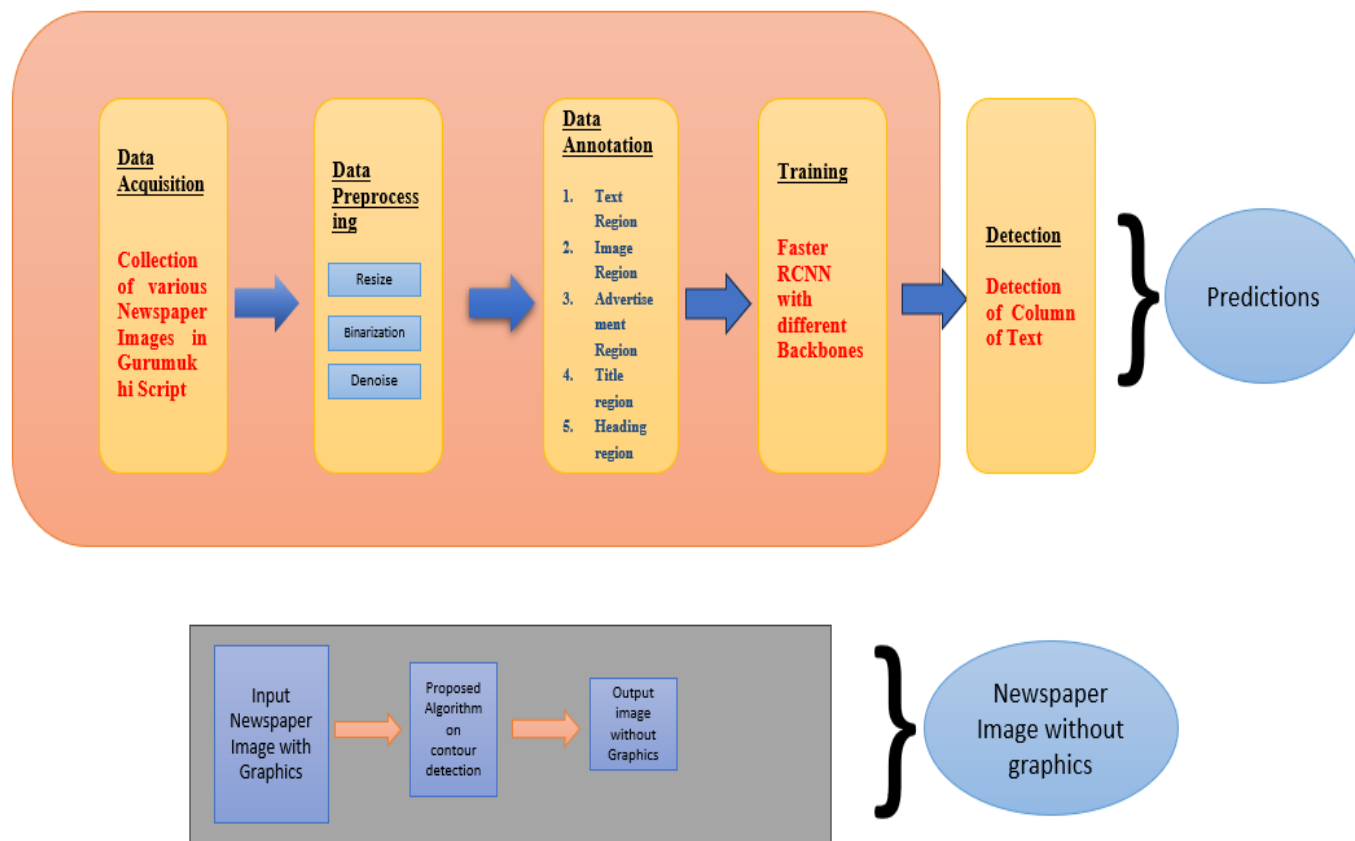


Figure 1. Workflow of the proposed system

al. (2007) proposed a technique to separate graphics by labelling components and combining components with similar features. Patil et al. (2022) created an approach based on pixel-based classification to find segment images from the text in documents. Alshameri et al. (2012) combined two algorithms—the run length smoothing algorithm and SVM algorithms with AND and OR to get a layout analysis of the image. Moreover, a dynamic horizontal projection is applied to the result based on the dynamic threshold. The benchmark dataset DocBank, which has 500K document pages and fine-grained annotated images for document layout analysis, was proposed by Li et al. (2020). Vasilopoulos and Kavallieratou (2017) used morphological methods and contour tracing to detect the complex layout of documents. Wu et al. (2023) proposed a dynamic fusion network based on features for layout analysis, giving an F1 score of 89.5% on the DSSE dataset and 95.1% on the CS-150. Kosaraju et al. (2019) proposed a multiclass classifier to segment documents into various components using a convolutional neural network. Bin Makhshen and Mahmoud (2019) surveyed various layout analysis techniques, features, advantages, and disadvantages. Biswas et al. (2021) used an object recognition mask RCNN to detect and predict regions of interest. Using straightforward image processing techniques like morphological dilation, run-length smoothing algorithm,

and rule-based algorithm, Soma and Shilpa (2022) offered a novel method to identify and segment blocks discovered within the newspaper regardless of layout. Zulfiqar et al. (2019) proposed a method to detect layout using the LSTM network, giving an accuracy of around 96%. Zhu et al. (2022) experimented with various algorithms to detect the layout of documents and created a dataset of 3000 newspaper pages in US states with various complex layout structures. Oliveira et al. (2018) proposed an open-source CNN architecture for layout segmentation with post-processing to improve accuracy. Xu et al. (2021) proposed a neural network built on the Mask R-CNN architecture to produce object classification, bounding box identification and the creation of page object masks. Alzubaidi et al. (2021) provided a detailed overview of various deep learning approaches, including CNN architectures and their uses and characteristics. Liebl and Burghardt (2021) evaluated various architectures for layout analysis of newspapers.

In this paper, we mainly proposed an algorithm to remove the graphics and pictures from newspapers in the Gurumukhi script. After that, we trained the newspaper images on a convolutional neural network by creating the dataset and segmenting the different columns of newspapers. The next parts of this paper mainly cover Material and Methods in Section 2, Results and Discussion in Section 3, and Conclusion in Section 4.



Figure 2. (a). Original Image (b) Result of the proposed algorithm after separating text from graphics (c) Original Image in Non-Manhattan layout (d) Result of the proposed algorithm after separating text from graphics

Materials and Methods

The workflow of the system has two parts. In the first part, the newspaper image is passed through an algorithm based on contours to remove the graphics from the newspaper images. In the second part, we collected around 500 images of Punjabi newspapers, performed data preprocessing and annotation into 5 classes, trained on Faster RCNN with different architectures, and performed inferences. The newspapers without graphics are then tested on the model to detect different columns of text, headline regions, and title regions. The workflow of the system is shown in Figure 1.

Proposed algorithm to Separate text from Images

The algorithm uses the concept of contour detection and performs various morphological operations to remove the graphics/photographs from the newspaper image. Firstly, newspaper images are binarized using various binarization algorithms (Sauvola and Pietikäinen, 2000). After the contours are detected, each contour's average height and width are obtained, and based on a predefined threshold, a newspaper image without graphics is generated.

Algorithm 1

1. Read the input image and convert the loaded image to grayscale.

2. Apply Sauvola thresholding to the grayscale image to create a binary image.
3. Erode the binary image using a kernel
4. Find contours in the eroded binary image
5. Initialize an empty list to store all contours.
6. Loop until all white pixels have been traced:
 - a. Find the first black (0) pixel as the starting point of a contour.
 - b. Initialize the current pixel as the starting point.
 - c. Initialize an empty list to store the points of the current contour.
 - d. Loop until returning to the starting point of the contour:
 - (i) Store the current pixel as a contour point.
 - (ii) Traverse in the current direction (right, down, left, up) to find the next black pixel.
 - (iii) Update the current pixel to the next pixel.
 - (iv) If the current pixel is the starting point, stop the loop.
 - (v) If the current pixel is the starting point, stop the loop.
 - e. Add the list of contour points (current_contour) to the list of all contours (all_contours).
7. Calculate the average height and width of all contours
8. Process each contour in the list of contours:



Figure 3. Labelling of Punjabi Newspaper Image using Labellmg Tool

- a. Get the center's bounding rectangle (x, y, width, height).
- b. If the height of the bounding rectangle is greater than the threshold times the average height:
 - i. Set the corresponding region in the original image to white.
10. Create a mask by drawing filled contours on a white canvas
11. Save the modified image and the mask to the specified paths.
12. Display the modified image without graphics.

In Figure 2, a is an original figure, and b is an image obtained by applying the proposed algorithm with text separated from graphics. Figures 2c and 2d also show that the proposed algorithm correctly removes pictures from the Manhattan layout as well as pictures from non-Manhattan layouts. We have tested the proposed algorithm on several images of complex layouts and found that it correctly separates text from pictures. Non-Manhattan layouts are those that form any arbitrary shape in newspapers.

Dataset Collection

In the second part of this paper, we first created the dataset. The dataset for newspaper layout analysis was

collected from various internet resources like websites and image resources. These newspapers are labeled into 5 classes Image_Region, Text_Region, Advertisement_Region, Heading_Region, and Title_Region with the help of the labelling tool Labeling as shown in Figure 3. Total number of Punjabi newspaper images collected is 500 and labelled. Table 1 represents a number of different regions annotated with the help of an annotation labeling tool. Total number of annotating objects is 20873.

Table 1. Distribution of Different labels in the dataset

Labels	Number of Annotations
Text_Region	8342
Image_Region	4563
Advertisement_Region	3421
Heading_Region	4332
Title_Region	215
Total	20873

Dataset Pre-Processing and splitting

As the images of newspapers are quite large, they are converted into a standard, smaller, and uniform size. The dataset is divided into training, testing, and validation. 70% of the images are in the training dataset, 20% in the testing dataset, and 10% in the validation dataset.

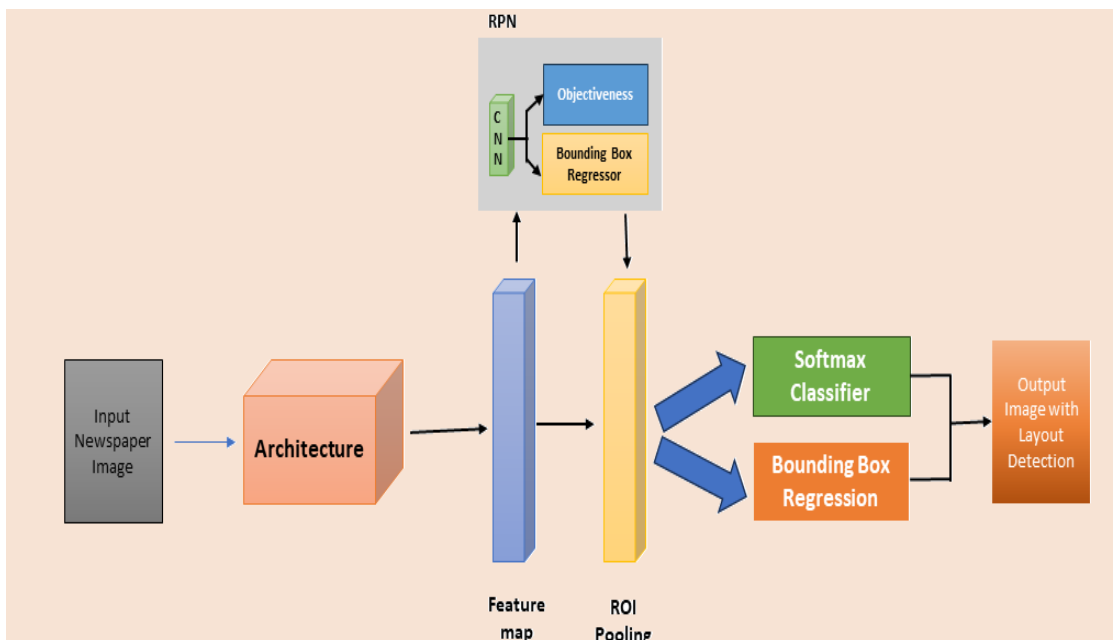


Figure 4. Faster RCNN Architecture

Deep Learning architectures

We have used an object detection model called Faster RCNN with different backbones: ResNET+FPN 50, 101, and Retinanet. Detectron2 which is a Facebook AI Computer Vision research framework that has various pre-trained models used for transfer learning. We have trained different architectures with the same configuration, and after that, testing is done on the test dataset. The images without graphics generated by the proposed algorithm are used for testing to detect the column of text.

Deep Learning Architecture Faster RCNN

Faster RCNN is made of CNN, the region proposal network, and prediction as shown in Figure 4. The RPN layer gives two outputs. One is objectiveness classification, and the other is a bounding box regressor, which acts as input to the ROI layer. The output of the ROI layer is given to the fully connected classification

layer, giving two outputs: a softmax classifier and a bounding box regression, giving the bounding box for each area of the newspaper image to detect the layout of the newspaper image.

Retinanet

RetinaNet typically uses a deep convolutional neural network (CNN) as its backbone. Popular choices include ResNet, ResNeXt, and other similar architectures. RetinaNet incorporates a feature pyramid network (FPN) on top of the backbone to handle objects of different sizes shown in Figure 5. RetinaNet employs anchor boxes at multiple scales and aspect ratios to propose potential object locations in the feature pyramid. These anchor boxes act as priors for object localization. RetinaNet employs separate classification and regression heads for each anchor. The classification head predicts the probability of each anchor containing an object of a specific class. The regression head predicts adjustments to

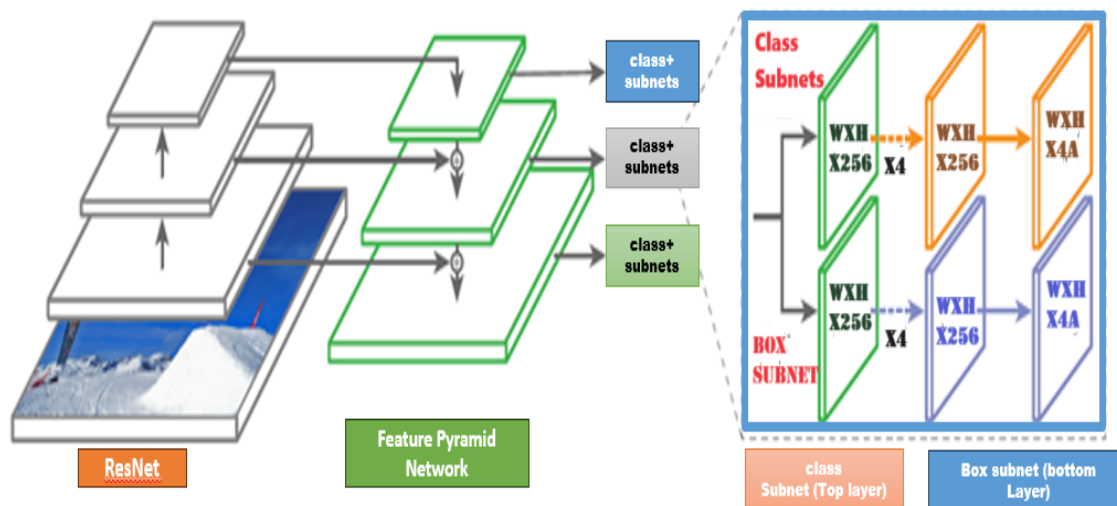


Figure 5. Architecture of Retinanet based on Resnet and Feature Pyramid Network to predict the Layout

the anchor box coordinates to fit the object's bounding box. The key innovation of RetinaNet is the focal loss function, designed to address the issue of class imbalance.

Results and Discussions

The experiments are conducted to remove the images of the Gurumukhi script from the newspapers based on the proposed algorithm. Around 100 newspaper images are tested using the proposed method. Table 2 shows the results and accuracy. We have trained the model based on Faster RCNN based on the configuration shown in Table 3.

Table 2. Performance of the proposed algorithm (Algorithm 1) to Separate text from Images

No of Newspapers	Total pictures in Newspaper images	Pictures removed	Accuracy
100	545	519	96.22

Table 3. Training System configuration of Convolution Neural network

Epoch: 300	Batch size: 4
Max Size Iteration: 300	Backbone: X101,R50
Optimizer: Default SGD with Detectron2	Number of classes: 6
Data Loader Worker: 4	Learning Decay: 0.001

Evaluation Criteria

1. **The IOU, or Jaccard Index**, is a metric that is the ratio of the area of intersection (the overlapping region) to the area of union (the combined region) of the predicted and ground truth regions used to test the performance.

$$IoU = (\text{Area of Intersection}) / (\text{Area of Union}) \quad (1)$$

2. **Precision:** Precision is a metric that measures the accuracy of positive predictions made by a model.

In the context of object detection, precision represents the fraction of the predicted bounding boxes that are correct. It is calculated as the ratio of true positives (correctly predicted objects) to the sum of true positives and false positives (incorrectly predicted objects).

$$\text{Precision} = (\text{True Positives}) / (\text{True Positives} + \text{False Positives}) \quad (2)$$

3. **Recall:** Recall, also known as sensitivity or true positive rate, measures the ability of a model to capture all positive instances. In object detection, it represents the fraction of the actual objects that the model correctly

detected. Recall is calculated as the ratio of true positives to the sum of true positives and false negatives (missed objects). $\text{Recall} = (\text{True Positives}) / (\text{True Positives} + \text{False Negatives}) \quad (3)$

4. **The F1-score** is another metric that combines precision and recall into a single value, which can be useful when you want to strike a balance between the two. It is calculated as the harmonic mean of precision and recall:

$$\text{F1-score} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

The different architectures are compared in Table 4 on the test dataset of newspaper images. The model's performance with X101 in this paper is better on mAP and F1_score, and the model's performance is more stable than other architectures. We have tested the model's performance on different architectures on 50 newspaper images. The accuracy of a model for layout detection using different architectures is shown in Table 5 and Figure 6.

Table 4. Comparison results of Faster RCNN with different architectures

Architecture	Precision	Recall	F1 Score
X101-FPN	0.87	0.81	0.838
R50-FPN	0.79	0.77	0.78
Retinanet R50	0.76	0.75	0.75
Retinanet101	0.81	0.83	0.82

Table 5. Performance of model with different architectures tested on 50 newspaper images

Model	Number of Text Columns	Detected Text Columns	Accuracy
X101-FPN	560	535	95.53
R50-FPN	560	521	93.03
Retinanet R50	560	502	89.64
Retinanet101	560	525	93.75

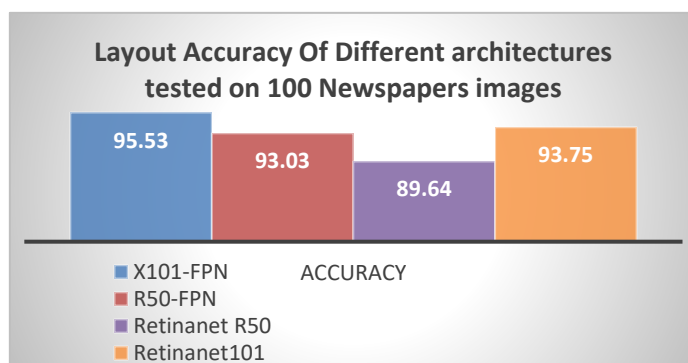


Figure 6. Validation Accuracy of different architectures for Layout Detection



Figure 7. Layout Detection Results on Punjabi Newspaper Image. Different blocks of text and the percentage of detection and label are shown



Figure 8. Layout Detection Results on Punjabi Newspaper Image. Different blocks of text and the percentage of detection and labels are shown.

We tested the model inference on different kinds of layouts by removing images or instances. In the case of X101-FPN, the model achieved an accuracy of 95.53%, which means it correctly detected 95.53% of the actual text columns. These metrics provide insights into the performance of the object detection models in identifying text columns within the dataset. A high accuracy percentage indicates that the model is proficient at correctly recognizing text columns, while lower accuracy percentages may suggest areas for improvement in the models or the dataset. The sample images detected are shown in Figures 7 and 8.

Conclusion

In the realm of newspaper digitization, particularly in the context of the Gurumukhi script, we presented a comprehensive approach to newspaper layout analysis for digitizing newspapers in the Gurumukhi script. This approach is critical to making historical newspapers accessible and user-friendly for modern audiences. We proposed an algorithm to effectively remove graphics and pictures from Punjabi newspaper images, achieving an accuracy of 96.22% on a test set of 100 images with diverse layouts, including non-Manhattan layouts. This algorithm also works on other newspaper scripts, like Hindi and English. The second part of the study focused on training a deep learning model, Faster RCNN, with different architectures to detect columns of text in newspapers. The dataset, consisting of 500 labeled Punjabi newspaper images, was used for training and testing. The model achieved high accuracy, with the best-performing architecture (X101-FPN) reaching an accuracy of 95.53% in detecting text columns. Evaluation metrics such as precision, recall and F1 score were employed to assess the performance of the deep learning models. The X101-FPN architecture demonstrated superior performance across these metrics, indicating its effectiveness in accurately identifying text columns within the newspapers.

The proposed methodology contributes to the field of newspaper digitization, particularly for languages like Gurumukhi, by addressing the challenges posed by complex layouts. The visual examples highlight the robustness of the proposed system for handling the complex layouts of Punjabi newspapers. The algorithm effectively removes graphics, and the deep learning model accurately detects different text regions within the newspapers. The qualitative assessment emphasizes the system's ability to handle diverse layouts, including non-manhattan, and showcases its potential for digitizing newspapers in the Gurumukhi script. The combination of

algorithmic techniques and deep learning models presented in this paper provides a robust solution for accurate layout detection, facilitating the digitization of historical newspapers for enhanced accessibility.

Future work could involve further refining the algorithm and model, exploring additional architectures, and expanding the dataset for improved generalization. Overall, this research lays the foundation for advancements in digitising newspapers in diverse scripts and layouts.

Conflict of Interest

The authors declare no conflict of interest.

References

- Alshameri, A., Abdou, S. M., & Mostafa, K. (2012). A combined algorithm for layout analysis of Arabic document images and text lines extraction. *International Journal of Computer Applications*, 49(23), 30–37. <https://doi.org/10.5120/7945-1282>
- Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A. Q., Duan, Y., Al-Shamma, O., Santamaria, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(53). <https://doi.org/10.1186/s40537-021-00444-8>
- BinMakhashen, G. M., & Mahmoud, S. A. (2019). Document Layout analysis. *ACM Computing Surveys*, 52(6), 1–36. <https://doi.org/10.1145/3355610>
- Biswas, S., Riba, P., Lladós, J., & Pal, U. (2021). Beyond document object detection: instance-level segmentation of complex layouts. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(3), 269–281. <https://doi.org/10.1007/s10032-021-00380-6>
- Chowdhury, S. P., Mandal, S., Das, A. K., & Chanda, B. (2007). Segmentation of Text and Graphics from Document Images. *International Conference on Document Analysis and Recognition ICDAR, Curitiba, Brazil*, 2, 619-623. <https://doi.org/10.1109/icdar.2007.4376989>
- Kosaraju, S., Masum, M., Tsaku, N. Z., Patel, P., Bayramoglu, T., Modgil, G., & Kang, M. (2019). DoT-Net: Document Layout Classification Using Texture-Based CNN. *2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia*, pp. 1029-1034. <https://doi.org/10.1109/icdar.2019.00168>

- Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., & Zhou, M. (2020). DocBank: A Benchmark Dataset for Document Layout Analysis. *arXiv*. <https://doi.org/10.18653/v1/2020.coling-main.82>
- Liebl, B., & Burghardt, M. (2021). An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers. *25th International Conference on Pattern Recognition (ICPR), Milan, Italy*, pp. 5153-5160. <https://doi.org/10.1109/icpr48806.2021.9412571>
- Oliveira, S. A., Séguin, B., & Kaplan, F. (2018). dhSegment: A Generic Deep-Learning Approach for Document Segmentation. *Arxiv*. <https://doi.org/10.1109/icfhr-2018.2018.00011>
- Patil, S., Vijayakumar, V., Mahadevkar, S., Athawade, R., Maheshwari, L., Kumbhare, S., Garg, Y., Dharrao, D., Kamat, P., & Kotecha, K. (2022). Enhancing Optical Character Recognition on Images with Mixed Text Using Semantic Segmentation. *Journal of Sensor and Actuator Networks*, 11(4), 63. <https://doi.org/10.3390/jsan11040063>
- Sauvola, J., & Pietikäinen, M. (2000). Adaptive document image binarization. *Pattern Recognition*, 33(2), 225–236. [https://doi.org/10.1016/s0031-3203\(99\)00055-2](https://doi.org/10.1016/s0031-3203(99)00055-2)
- Soma, S., & Shilpa. (2022). A Novel Approach for Newspaper Block Segmentation using Run-Length Smoothing Algorithm. *2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India*, pp. 1-6. <https://doi.org/10.1109/icmnwc56175.2022.10031933>
- Vasilopoulos, N., & Kavallieratou, E. (2017). Complex layout analysis based on contour classification and morphological operations. *Engineering Applications of Artificial Intelligence*, 65, 220–229. <https://doi.org/10.1016/j.engappai.2017.08.002>
- Wu, X., Ma, T., Du, X., Hu, Z., Yang, J., & He, L. (2023). DRFN: A unified framework for complex document layout analysis. *Information Processing and Management*, 60(3), 103339. <https://doi.org/10.1016/j.ipm.2023.103339>
- Xu, C., Shi, C., Bi, H., Liu, C., Yuan, Y., Guo, H., & Chen, Y. (2021). A page object detection method based on mask R-CNN. *IEEE Access*, 9, 143448–143457. <https://doi.org/10.1109/access.2021.3121152>
- 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>
- Zulfiqar, A., Ul-Hasan, A., & Shafait, F. (2019). Logical Layout Analysis using Deep Learning. *Digital Image Computing: Techniques and Applications (DICTA), Perth, WA, Australia*, 1-5. <https://doi.org/10.1109/dicta47822.2019.8946046>

How to cite this Article:

Atul Kumar and Gurpreet Singh Lehal (2023). A Hybrid Approach for Complex Layout Detection of Newspapers in Gurumukhi Script Using Deep Learning. *International Journal of Experimental Research and Review*, 35(Spl.), 34-42.

DOI : <https://doi.org/10.52756/ijerr.2023.v35spl.004>



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.