IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Deep Learning-Based Detection of One and Two-Column Textual Blocks in Camera-Captured Pashto Documents Images

BADSHAH SAEED[1], SIRAJ MUHAMMAD[1], KHALIL ULLAH[2], ABDALLAH NAMOUN[3], AHMAD KHAN[4] IKRAM SYED [5], SYED SAJID ULLAH[6], IBRAR HUSSAIN[7]

[1]Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Upper Dir, KP, Pakistan (e-mail:saeed@sbbu.edu.pk, msiraj83@sbbu.edu.pk)
[2]Department of Software Engineering, University of Malakand (UOM), Khyber Pakhtunkhawa (KP), Pakistan (e-mail: khalil.ullah@uom.edu.pk)
[3]Faculty of Computer Science and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia (e-mail: a.namoun@iu.edu.sa)
[4]Faculty of Computer Science, The Superior University, Lahore, (Faisal Abad Campus), Faisal Abad, Punjab (e-mail: ahmadkhan46@hotmail.com)
[5]Dept Information & Communication Engineering, Hankuk University of Foreign Studies, Yongin 17035, South Korea (e-mail: ikram@hufs.ac.kr)
[6]Department of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway (e-mail: syed.s.ullah@uia.no)
[7]QEC, Shaheed Benazir Bhutto University, Sheringal, Dir Upper (e-mail: ibrar@sbbu.edu.pk)

Corresponding author: Ibrar Hussain, and Abdallah Namoun (e-mail: ibrar@sbbu.edu.pk).

ABSTRACT

The paper explores the layout analysis and classification task of Pashto document images, a field with limited research due to the language's low-resource status. It uses Document Image Analysis (DIA) to detect one-column and two-column text blocks from Pashto documents captured using handheld cameras. A novel dataset containing real-world documents is annotated using bounding boxes to distinguish between one-column and two-column layouts. Layout analysis plays a vital role in Document Image Analysis (DIA) and answers questions about the nature of Document Images integral parts (components). These components are mainly text, graphics, and tabular data. The detection/ identification of such elements is initially dependent on the general layout of a document. Textual columns are one of the basic layout structures present in document images. Such textual columns need identification as one-column or two-column before feeding the textual blocks into an Optical Character Recognition (OCR) system. Also, it becomes more crucial if the document images belong to a low-resource language like Pashto. Thus, this work contributes mainly in two aspects. First, it contributes to creating a dataset containing camera-captured document images with one-column and two-column patterns. Second, it proposes a deep learning model to identify/ detect one-column and two-column textual blocks in the Pashto text images. The proposed model utilizes a reputed variant of a Faster Region-based Convolutional Neural Network (R-CNN) called single shot detector (SSD). The evaluation was done by examining the test set and a mean average precision (mAP) of 89% is achieved as a baseline.

INDEX TERMS DIA, Deep Learning Models, Pashto, SSD

I. INTRODUCTION

The automatic detection and analysis of textual blocks in document images has emerged as a vital issue in document image analysis (DIA) systems, particularly for languages with right-to-left scripts like Pashto. Significant development has been made in text detection for Latin-based scripts (1). Languages like Pashto present precise challenges due to their unique character set, similar-shaped characters, dots and diacritics, cursive nature, and complexities of the writing system (2). Document images are those digital images that are achieved either from a scanner or a camera. These documents include articles, postal addresses, bank cheques, forms, topographic maps, engineering drawings, license plates, and billboards etc. (3). The main sources for the acquisition of document images are scanners and cameras. These document images are in pixel form, and they could not be searched and analyzed in computers (4). In addition to that, such images occupy large space on a computer's storage and hence present a chal-

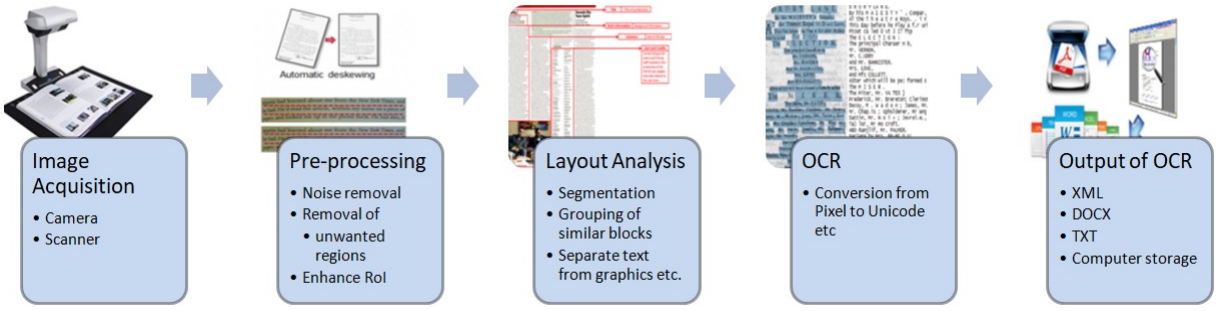


FIGURE 1: A generic overview of a DIA system

lenge to the space factor. Camera-captured document images introduce additional complexities compared to conventional scanned documents, including varying perspective distortions, illumination conditions, and blur effects (5). These complexities are mainly stated when handling documents containing multiple column layouts, which are standard in Pashto academic papers, newspapers, and books (6).

Reading the text from the document images captured by the camera is an important process in different applications such as digitization and retrieval. However, it is often challenging because of factors such as Scene complexity/variation in the lighting conditions and Different language scripts. In the context of the current research paper, we focus on detecting one and two-block text regions from Pashto document images captured through a camera. To solve and convert such document images into a digital or readable format, we need a specialized technique/ research area where we could analyze document images. Thus, such an area of research is named Document Image Analysis or shortly (DIA).

DIA is an area of research that comprises algorithms, techniques, and processing steps that are used to obtain information that a computer can understand from pixel-sized document images (3; 7; 8; 9). The major components in a generic DIA system are (1) Image Acquisition (2) Preprocessing, (3) Layout Analysis and Classification, (4) Optical Character Recognition (OCR), and (5) OCR's Output. Fig 1 illustrates a typical diagram of a general DIA system. However, this research examines and contributes mainly to the Layout out Analysis step.

Further, Pashto is a low-resource language and has very little work regarding the DIA system (10; 11; 12)that addresses the recognition of textual blocks via text-line images. However, considering the overall documents containing Pashto language have not been explored in terms of Layout Analysis. Therefore, this research examines and explores the material of Pashto documents in terms of layout analysis focusing on the detection of one-column and two-column textual blocks.

Pashto is an Indo-Iranian language by spoken in Afghanistan and northwestern Pakistan; It has an Arabic based script like Persian. Pashto script has certain features that differ from those of English script, coupled with the fact that camera-captured images of documents are not perfectly clear and problematically contain both background and foreground information, this makes the job of accurate localization and segmentation of text regions in the Pashto script from document images very challenging (6; 13).

In (14) and (15) the authors work on Arabic verse Pashto text and graphics verse text detection. However, their contribution did not cover one column vs two-column detection. Therefor, in this paper, we have suggested a deep learning method for one and two column text block detection in the Pashto document images. There are two major contributions to this research. The first contribution is the creation of a dataset that contains camera-captured Pashto documents with one-column and two-column textual patterns. Further, to support supervised learning, the dataset is well transcribed by selecting the one-column and two-column textual blocks with closed bounding boxes. The second contribution is the development of a deep learning-based classifier that can detect one-column and two-column textual blocks as separate classes. However, the application of these algorithms to Pashto documents is still largely unexplored, especially in the context of camera-captured images with varied column layouts. Our proposed classifier is based on the utilization of Single Shot Detector (SSD) (16) which is a variant of Faster Region-based Convolutional Neural Network (Faster R-CNN) (17). The results show 89% accuracy as a baseline on the newly created dataset.

This study proposes a novel dataset and a deep learning-based approach for automatically detecting and classifying one and two-column textual blocks in camera-captured Pashto document images. Our method addresses the challenges of Pashto text processing and maintains robustness to the diverse distortions found in camera-captured documents. The proposed system aims to facilitate more efficient digitization and analysis of Pashto documents, contributing to the preservation and accessibility of content in this critical language. The main contribution of this work is the Camera Captured Pashto Printed Text Imagebase (CCPPTI) and the baseline on SSD and YELO11, for one and two-column textual block detection.

II. RELATED WORK

The detection and analysis of textual blocks in document images has advanced dramatically in the last decade, notably with the introduction of deep learning algorithms. Significant work regarding the area of DIA systems has grown more rapidly in the late 1980s and early 1990's (18). However, to be more precise, we only report the related work that is more focused on layout analysis and classification in the field of DIA. For simplicity purposes, we have split and narrowed down the related work into important subfields: layout analysis and classification.

A. RELATED WORK REGARDING LAYOUT ANALYSIS

L O'Gorman et al. (7) described the document spectrum, or docstrum, for structural page layout analysis. The base of the developed method was on bottom-up approach and the clustering of the nearest- neighbor components of page. The method produced results on an accurate measurement of skew, within-line and between line spacing's. Besides,the method also locates the text lines and text blocks.

Simon et al. (19) described a new method for document layout analysis that is based on bottom-up approach. The implementation of the method done in the CLiDE[1] and based on Kruskal's algorithm. The method is appropriate for a broad range of documents. This method is applied to make the physical appearance of the page by using a defined distance metric between the page components. Major advantages of their method are speed and generality. The achieved good accuracy of the method by analyzing 98 test images.The method reported rate of error of about 1%.

Breuel et al. (20) introduced some novel algorithms and statistical methods for layout analysis. They have evaluated their system/approach on a subset of UW3[2] database. Their algorithm yields better performance regarding layout analysis.

Kevin Laven et al. (21) investigated the effectiveness of statistical pattern recognition algorithms. They used the developed algorithm for logical and physical layout analysis issues. They solved these problem by following many rule based, grammar based methods. They have created a data set [3] of 932 page images from different academic journals. They achieved average precision accuracy rate of 85.5% among 16 categories and 86.0% generalized average accuracy rate for unidentified number of categories.

A. Antonacopoulos et al. (22) discussed the ground truth performance evaluation of layout analysis. They also discussed different issues related to the ground truth surroundings of Layout analysis methods. Their developed data-set [4] has been made freely available to researchers. Their achievements in this field have been selected for two international competitions in international conferences held in 2005 and 2007.

Shafait et al.(23) followed the Breuel model for page layout analysis. They introduced a text line model for converting Urdu text lines into Nastaliq script. For performance evaluation of the model, they scanned Twenty Five Urdu documents collected from several different sources. These sources were classified into five(05) classes i.e. book, poetry, digest, magazine, newspaper. In the data-set, there are 5 images of each class. Their dataset[5] has been made publicly available. The developed system was tested with data of selected classes, and 92% accuracy was achieved for book, magazine and poetry documents due large inter line spacing. Though the accuracy of digest decreases to 80% due to small spacing between the lines and also due to different enumerated lists present in the document. For newspaper the accuracy falls to 72% because of several fonts sizes, inverted texts, small inter-line spacing and the low quality of the page.

Ray Smith et al. (24) developed a new algorithm of hybrid nature for page layout analysis. To form the initial data type hypothesis they used a bottom-up methods. They used tab-stops to impose structure and reading order on detecting regions. The achieved an accuracy rate of 92% for the developed method.

M. Sezer Erkilinc et al. (25) introduced an algorithm for document classification and layout analysis. They evaluate the developed method by different samples of document images having simple, complex nature along with other characteristics of color and gray-scale. They used two modules in their system. The first module is used for detection of text. This module has two techniques RLE[6] (26) and Hough transform and edge linkage analysis. The second module is used for lines and strong edges detection. They achieved an average accuracy rate of 85%.

Tuan-Anh Tran et al.(27) proposes a method for the textual and non-textual classification of document images. They used a combination of two methods whitespace analysis and multi-layes homogenous regions. This combination is known as recursive filters. Their method superiority and effectiveness have been proved from the experimental results obtained on ICDAR 2009 competition for dataset of page segmentation. They achieved an average 90% above accuracy.

R Ahmad et al.(28) developed a method for the extraction of text-line in Arabic script. Their methods extracted large headings and titles in Arabic script. Their method is based on two techniques i.e. Horizontal projection profile (HPP)(29) and Hanning window smoothing technique.They achieved an accuracy rate of 99.30%. This method is limited to de-skewed images.

Deep learning algorithms have significantly improved multi-column layout identification. The study of (30) suggested a Mask R-CNN-based technique that detected numerous columns with 94.3% accuracy in English documents.

[1](Chemical Literature Data Extraction) system (http://chem.leeds.ac.uk/ICAMS/CliDE.html)

[2]University of Washington Database3

[3]http://jmlr.csail.mit.edu

[4]:http://www.prima.cse.salford.ac.uk/dataset

[5]http://www.iupr.org/demos downloads/

[6]Run Length Encoding

However, their method needed major modification for right-to-left scripts. More recently, (14) created a YOLOv5-based architecture tailored exclusively for multi-script documents, displaying significant performance in Arabic vs Pashto text detection in camera-capture document images.

B. DATASET CREATION

The dataset consists of camera-captured document images featuring Pashto text in one or two columns. We collected images from real-world documents with varying layouts, lighting, and font sizes. Deep learning models can train on Pashto text layouts, which typically have one-column and two-column structures, for layout recognition. The images are acquired from periodical, books, journals, and magazines consisting of one and two-column text.The name of the dataset is Camera-Capture Pashto Printed Text Imagebase (CCPPTI). The main objective of this research is to classify one-column and two-column textual data present in Pashto text images, it is not possible to achieve such classification without appropriate data. Therefore, the creation of a dataset becomes another important milestone for this research. Dataset creation is always a laborious task and requires scientific skills to create it. Thus, we create a real-world-based dataset that will contain camera-captured document images. The images will contain mainly Pashto text using a pattern of one-column and or two-column. Each image will have its corresponding ground-truth information in a separate file. The ground truth information is labeled using LabelMe (open source software). We use pixel-level labeling, that will mask the entire textual blocks either one-column or two-column. Fig 2 shows the input image and its corresponding visual illustration via ground truth.

III. PROPOSED METHODOLOGY

Our method involved using a Convolutional Neural Network (CNN) in conjunction with a Single Shot Detector (SSD) to accomplish a reliable layout analysis of Pashto document images. The SSD model, well-known for its quickness and effectiveness in object detection, allows for both class identification and bounding box prediction to be done simultaneously using feature maps. SSD predicts boundary boxes and class scores from feature maps in a single pass, eliminating the need for a delegated region proposal network, in contrast to its predecessor, the Region-based Convolutional Neural Network (R-CNN). With VGG16 as its feature extraction backbone, this network is especially well-suited to object detection issues and provides a simplified method for real-time detection jobs. Multiple convolutional layers intended to learn unique properties at different scales make up the SSD architecture. Like traditional CNN filters, SSD uses VGG16 layers to extract feature maps, which are subjected to additional 3x3 convolutional filtering at each cell to generate predictions. This method works effectively with Pashto text graphics, where text alignment and spatial arrangements can vary greatly. To achieve the objectives and to explore the layout analysis phase of the DIA system in the Pashto language, we will use the Convolution Neural Network (CNN) with a deep learning approach. For the subject purpose, we will choose mainly Single Shot Detector (SSD) (16) and Yolov11 (31).

A. SSD

Unlike its predecessor R-CNN(17), the SSD does not use a delegated region proposal network, instead, it predicts the boundary boxes for the classes directly from feature maps in a single pass. This network is famous for its speed and performance regarding object detection problems. The architecture of SSD mainly comprises VGG16 layers, convolutional layers, and feature maps. SSD uses VGG16[7] to extract feature maps, then it detects objects using the convolution in VGG16 layers (32). The SSD architecture is trained on more than a million images of the ImageNet database [8]

It computes both the location and class scores using small convolutional filters. After extracting the feature-maps, SSD applies 3×3 Convolutional filters for each cell to make predictions. These filters compute the results just like the regular CNN filters. The typical SSD model is shown in Figure 3.

B. YOLOV11

Building on the success of earlier YOLO versions, YOLOv11 is a state-of-the-art (SOTA) model that adds new features and enhancements to increase performance and flexibility further. Because of its quick, accurate, and user-friendly architecture, YOLOv11 is an excellent option for a variety of object recognition and tracking, instance segmentation, image classification, and posture estimation tasks (34; 35).

C. EVALUATION CRITERIA

We will use mAP (mean average precision) and IoU (intersection over union) for the evaluation of our proposed model. The mentioned metrics are briefly explained below.

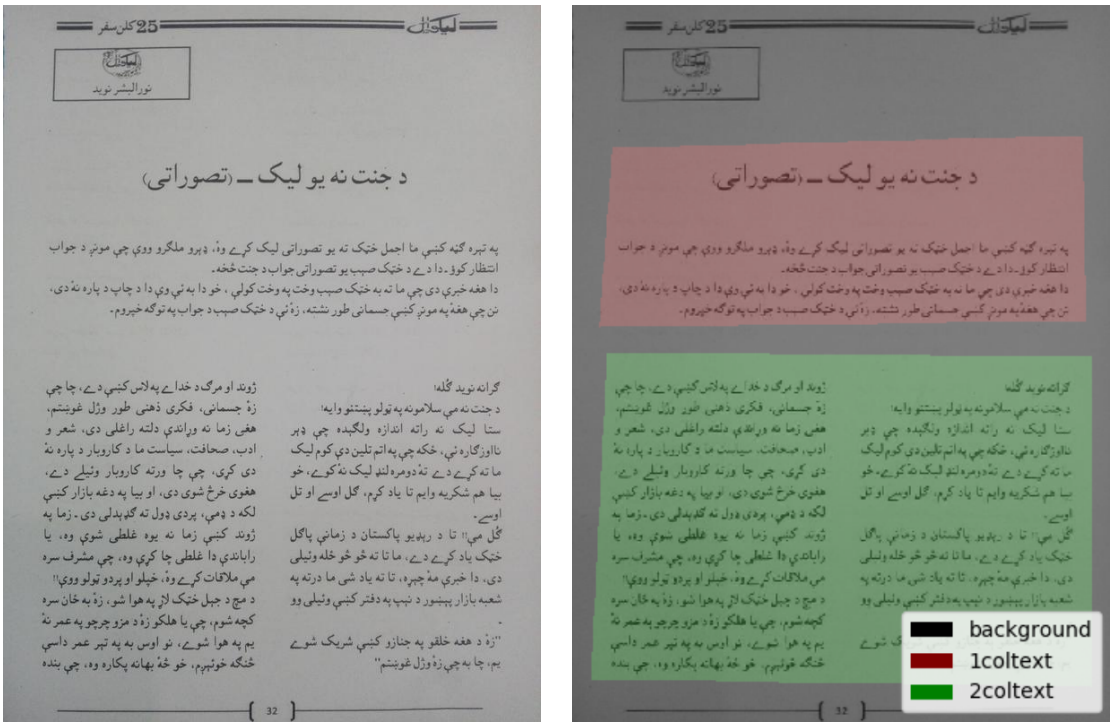
- AP (Average precision): AP is a popular metric in measuring the accuracy of object detectors like Faster R-CNN, SSD, etc. Average precision computes the average precision value for recall value over 0 to 1(36). While, Mean average precision for a set of queries is the mean of the AP scores for each query. The mAP can be calculated via equation 1.

$$mAP = \frac{\sum_{q=1}^{Q} avP(q)}{Q} \tag{1}$$

  Where avP(q) is the average precision (AP) for a given query and Q is the total number of queries.
- IoU (Intersection over union): IoU measures the overlap between 2 boundaries. IoU is used to measure how much the predicted boundary overlaps with the ground

[7]VGG16 is the creation of Visual Geometry Group (VGG), and contains 16 convolutional layers.

[8]The ImageNet project is a large visual database designed for the recognition of visual objects. The database consists of more than 14 million images (33).



(a) (b)

FIGURE 2: Input image (a), and (b) is the corresponding ground truth

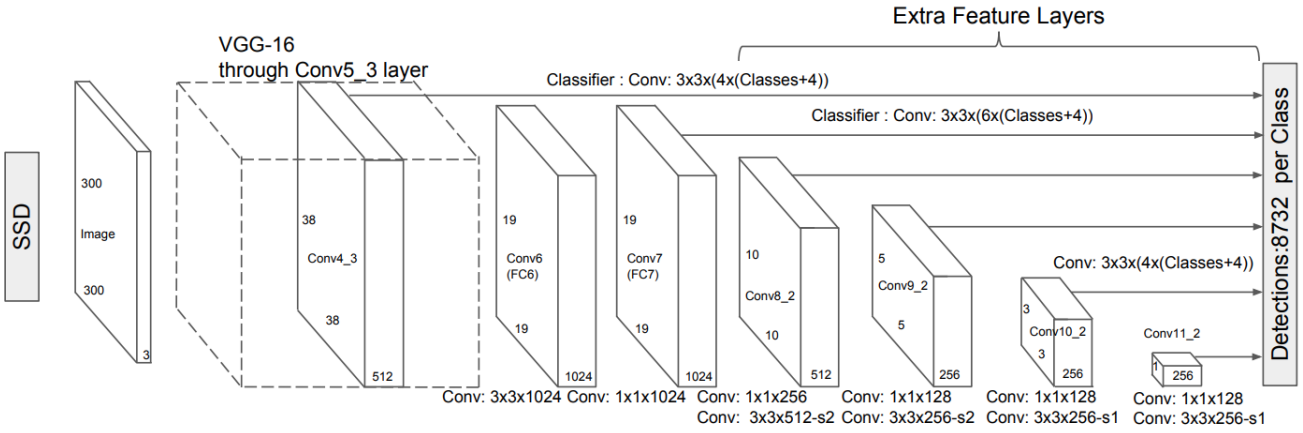


FIGURE 3: A Typical Single shot detector (SSD) Model (16)

truth bounding-box(the real object boundary)(36). IoU can be measured via equation 2.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{2}$$

Model	SSD	Yolov11
mAP @ IoU =0.50	89%	94%

TABLE 1: Mean average precision for SSD and Yolov11 using Intersection over union (0.50)

IV. RESULT AND DISCUSSION

A. RESULTS

After 29k epochs, we have stopped the training, and the final model has selected for evaluation. The evaluation was done by examining 330 images from the test set. In this way, a mean average precision of (mAP) 89% has been achieved while in case of Yolov11 94% were achived. Figure 4 shows the mAP in detail along with corresponding epochs as well as total elapsed time.

Similarly, the training process was tested for how it reduces the total loss. The total loss that we finally achieved is 3.6, while its ceilings was approximately 5. Figure **??** depicts

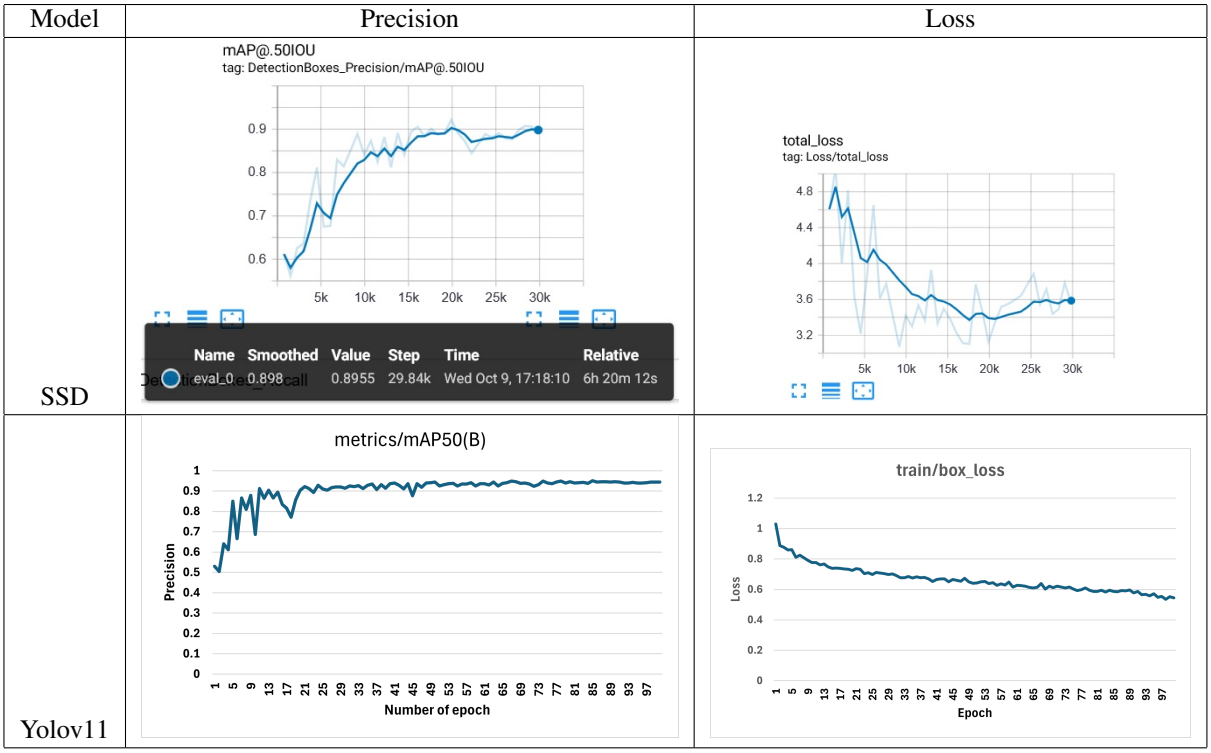


FIGURE 4: Training loss and precision for SSD and Yolov11

the value of loss w.r.t epochs in the training.

B. DISCUSSION

After rigorous analysis of results as well as individual images we have major findings that are explained in the following text. Before, we could empirically asses the results, FIGURE 5 and FIGURE 6 shows some prediction of our proposed model on completely unseen data. The light green color represent the bounding box for "onecolumntext" and light blue color represent "twocolumntext" for SSD while the light green color represents "onecolumntex" and dark blue color represents "twocolumntex" for Yolov11 model. FIGURE 5 shows that our model has produces good performance. The images shows the 99% prediction for both "onecolumntext" and "twocolumntext" for SDD model. It can be clearly seen that near to 99% prediction of true positive classes are achieved. However, in case of Yolov11 the predicted bounding boxes are more accurate in term of text enclosed. Moreover, FIGURE 5 shows the tested images with visual illustration of higher precision.

On the other hand FIGURE 6 shows miss-classification and low detection.

The first row of FIGURE 6 images shows the 88% average detection for SSD model. The Second and third row of FIGURE 6 show that SSD model miss some text blocks for onecolumntext as well twocolumntext while Yolov11 correctly classify that missing text block. This shows that in general Yolov11 shows better results than SSD model.

It is concluded that for small bounding-boxes the predictions is not good as for large bounding-boxes. Further investigation produced that the examples for the small textual blocks are approximately less than the large textual block. This is changeable distribution guide the classifier to gain an understand about large textual block rather then small textual block. Further validated this point explicitly and examine the mAP for only small bounding-boxes, and hence it is proved that the mAP for small bounding box is just 25%.

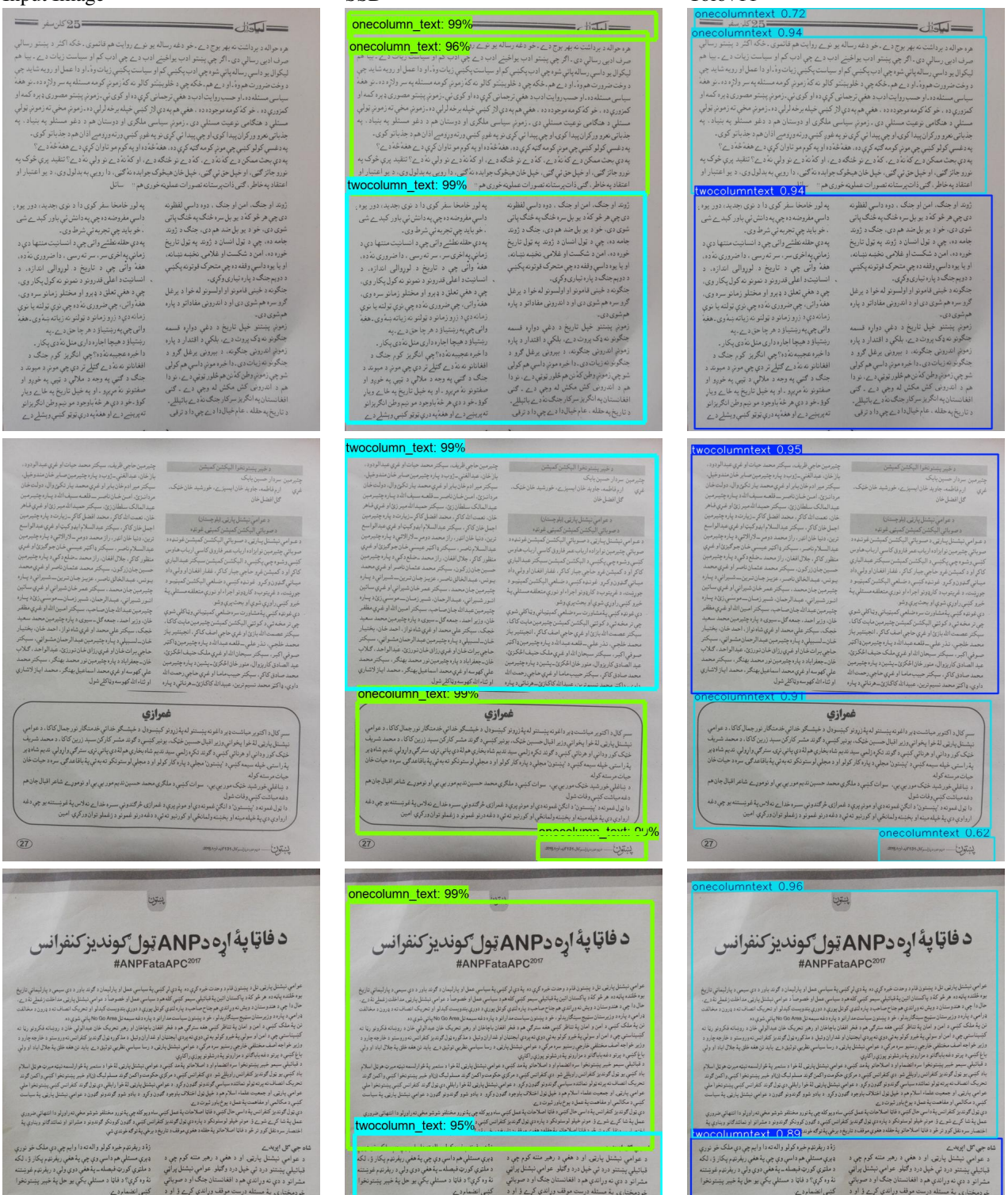


FIGURE 5: Tested images and their visual illustration with high precision.

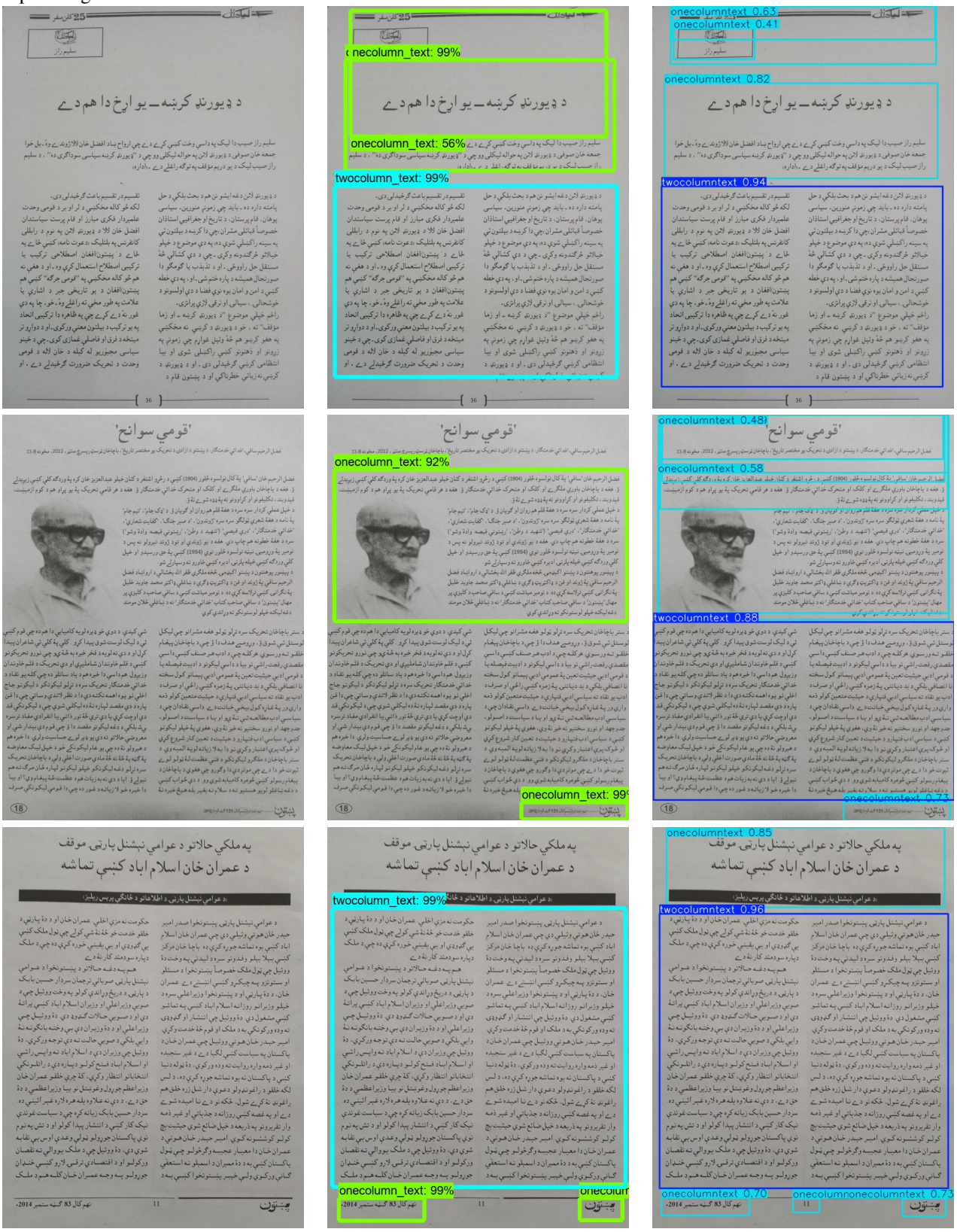


FIGURE 6: Tested images and their visual illustration with high precision.

V. CONCLUSIONS, LIMITATIONS AND FUTURE WORK

A. CONCLUSION

This study for the first time presents a pioneering study regarding the layout analysis and classification of Pashto document images. The research particularly examined the classification of one-column text and two-column text in Pashto document images. For this purpose, we have first created a new dataset that contains real Pashto document images. The images are acquired via a handheld camera. The dataset is freely available and will be a significant resource for the research community for analyzing the DIA domain in cursive scripts. Further, this research goes one step ahead and applies the deep learning-based method to examine how we can detect/ classify Arabic text and Pashto text in a single document image. We have chosen the SSD model that has a hybrid model containing VGG16 as convolutional layers and a Neural network for learning highly distinctive features. Regarding the Pashto language, this work is the first work, that uses an architecture based on deep learning and examines the Layout and classification stage of a DIA system. We have achieved the overall mAP of 89% as a benchmark. The results are promising as it is the first-ever attempt to analyze and investigate the Pashto language w.r.t the DIA system.

B. LIMITATIONS

The research shows promising results but has several limitations that could affect their generalization and accuracy. The small bounding boxes used to label textual blocks yielded suboptimal results, as they may not capture the irregular contours of the Pashto script, especially within a single image. The dataset's limited instances of small textual blocks also affected the classifier's ability to generalize across text sizes and formats. Further dataset expansion and improved bounding box methods are needed to achieve finer granularity and better classify smaller text blocks. The model architecture, designed to recognize rectangular regions, struggles with irregularly shaped text, leading to less precise localization and classification.

C. FUTURE WORK

While analyzing and discussing the results we have found two major points:

1) The textual blocks which are comparatively small or bounded in a small bounding-box show poor results. We need further research to empirically investigate this issue and find out the real causes that lead us to such poor results. The reasons could be the small size of the dataset or the smaller number of instances specifically related to small bounding boxes.
2) We have used rectangular bounding boxes to make ground-truth information for training. The rectangular bounding boxes are very suitable for regular shapes while textual blocks are mainly irregular. In short, polygon-based labeling will be another dimension to explore as a future work.
3) Further research could explore the impact of neural network architecture modifications on detection accuracy, potentially leading to more accurate and computationally efficient models.
4) The model's performance in real-world applications could be enhanced by incorporating context-aware techniques that analyze spatial relationships and layout patterns in documents.
5) This research focuses on Pashto script, but future studies could explore adaptability to other languages and scripts, enhancing the inclusive and utility of Document Image Analysis systems globally.

Other dimensions might be about to change the configuration of the Neural Networks. For example the size of the layers and its impact on accuracy can be empirically examined in future.

REFERENCES

[1] M. Shabana, A. Jose, and A. Sunny, "Text detection and recognition in natural images," International Research Journal of Engineering and Technology, vol. 05, pp. 995–999, 2018.
[2] I. Hussain, R. Ahmad, K. Ullah, S. Muhammad, R. Elhassan, and I. Syed, "Deep learning-based recognition system for pashto handwritten text: benchmark on phti," PeerJ Computer Science, vol. 10, p. e1925, 2024.
[3] K. D. Kalaskar and M. P. Dhore, "Preprocessing challenges in document image analysis," in Preprocessing challenges in document image analysis.(MPGINMC, 2012, pp. 1–4.
[4] U. D. Dixit and M. Shirdhonkar, "A survey on document image analysis and retrieval system," International Journal on Cybernetics & Informatics (IJCI), vol. 4, no. 2, pp. 259–270, 2015.
[5] D. Doermann, J. Liang, and H. Li, "Progress in camera-based document image analysis," in Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings. IEEE, 2003, pp. 606–616.
[6] I. Hussain, R. Ahmad, S. Muhammad, K. Ullah, H. Shah, and A. Namoun, "Phti: Pashto handwritten text imagebase for deep learning applications," IEEE Access, vol. 10, pp. 113 149–113 157, 2022.
[7] L. O'Gorman and R. Kasturi, Document image analysis. IEEE Computer Society Press Los Alamitos, 1995, vol. 39.
[8] D. Salvi, "Document image analysis techniques for handwritten text segmentation, document image rectification and digital collation," Ph.D. dissertation, University of South Carolina, 2014.
[9] E. Balamurugan, K. Sangeetha, and P. Sengottuvelan, "Document image analysis - a review," 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:64999069
[10] R. Ahmad, S. Naz, M. Z. Afzal, S. H. Amin, and T. Breuel, "Robust optical recognition of cursive pashto

script using scale, rotation and location invariant approach," PloS one, vol. 10, no. 9, p. e0133648, 2015.

[11] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, T. Breuel, and A. Dengel, "Kpti: Katib's pashto text imagebase and deep learning benchmark," in 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). IEEE, 2016, pp. 453–458.

[12] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, and T. Breuel, "Scale and rotation invariant ocr for pashto cursive script using mdlstm network," in 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2015, pp. 1101–1105.

[13] M. Ataullha, M. H. Rabby, M. Rahman, and T. B. Azam, "Bengali document layout analysis with detectron2," arXiv preprint arXiv:2308.13769, 2023.

[14] N. Khan, R. Ahmad, K. Ullah, S. Muhammad, I. Hussain, A. Khan, Y. Y. Ghadi, and H. G. Mohamed, "Robust arabic and pashto text detection in camera-captured documents using deep learning techniques," IEEE Access, vol. 11, pp. 135 788–135 796, 2023.

[15] K. Bahadar, R. Ahmad, K. Aurangzeb, S. Muhammad, K. Ullah, I. Hussain, I. Syed, and M. S. Anwar, "Pashto script and graphics detection in camera captured pashto document images using deep learning model," PeerJ Computer Science, vol. 10, p. e2089, 2024.

[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in European conference on computer vision. Springer, 2016, pp. 21–37.

[17] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[18] P. M. Manwatkar and K. R. Singh, "A technical review on text recognition from images," in Intelligent Systems and Control (ISCO), 2015 IEEE 9th International Conference on. IEEE, 2015, pp. 1–5.

[19] A. Simon, J.-C. Pret, and A. P. Johnson, "A fast algorithm for bottom-up document layout analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 3, pp. 273–277, 1997.

[20] T. M. Breuel, "High performance document layout analysis," in Proceedings of the Symposium on Document Image Understanding Technology, 2003, pp. 209–218.

[21] K. Laven, S. Leishman, and S. Roweis, "A statistical learning approach to document image analysis," in Eighth International Conference on Document Analysis and Recognition (ICDAR'05). IEEE, 2005, pp. 357–361.

[22] A. Antonacopoulos, D. Karatzas, and D. Bridson, "Ground truth for layout analysis performance evaluation," in International Workshop on Document Analysis Systems. Springer, 2006, pp. 302–311.

[23] F. Shafait, "Geometric layout analysis of scanned documents," Ph.D. dissertation, Technische Universität Kaiserslautern, 2008.

[24] R. W. Smith, "Hybrid page layout analysis via tab-stop detection," in 2009 10th International Conference on Document Analysis and Recognition. IEEE, 2009, pp. 241–245.

[25] M. S. Erkilinc, M. Jaber, E. Saber, P. Bauer, and D. Depalov, "Page layout analysis and classification for complex scanned documents," in Applications of Digital Image Processing XXXIV, vol. 8135. International Society for Optics and Photonics, 2011, p. 813507.

[26] D.-H. Xu, A. S. Kurani, J. D. Furst, and D. S. Raicu, "Run-length encoding for volumetric texture," Heart, vol. 27, no. 25, pp. 452–458, 2004.

[27] T.-A. Tran, I.-S. Na, and S.-H. Kim, "Separation of text and non-text in document layout analysis using a recursive filter." KSII Transactions on Internet & Information Systems, vol. 9, no. 10, 2015.

[28] R. Ahmad, M. Z. Afzal, S. F. Rashid, M. Liwicki, and A. Dengel, "Text-line segmentation of large titles and headings in arabic like script," in 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). IEEE, 2017, pp. 168–172.

[29] M. Javed, P. Nagabhushan, and B. Chaudhuri, "Extraction of projection profile, run-histogram and entropy features straight from run-length compressed text-documents," arXiv preprint arXiv:1404.0627, 2014.

[30] A. Almutairi and M. Almashan, "Instance segmentation of newspaper elements using mask r-cnn," in 2019 18th IEEE International conference on machine learning and applications (ICMLA). IEEE, 2019, pp. 1371–1375.

[31] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[32] A. Canziani, A. Paszke, and E. Culurciello, "An analysis of deep neural network models for practical applications," arXiv preprint arXiv:1605.07678, 2016.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[34] S. Ullah, R. Ahmad, A. Namoun, S. Muhammad, K. Ullah, I. Hussain, and I. A. Ibrahim, "A deep learning-based approach for part of speech (pos) tagging in the pashto language," IEEE Access, 2024.

[35] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," arXiv preprint arXiv:2410.17725, 2024.

[36] M. Everingham and J. Winn, "The pascal visual object classes challenge 2012 (voc2012) development kit," Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep, 2011.

BADSHAH SAEED did his BS (Hons) from the Department of Computer Science, University of Malakand, Pakistan and MS (Computer Science) from the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Dir Upper.

SIRAJ MUHAMMAD Completed his MPhil from Quaid-i-Azam University, Islamabad in 2010 and did his PhD in 2020 from Asian Institute of Technology (AIT), Thailand. He was a Software Engineer in Elixir Technologies of Pakistan, Islamabad from 2010 to 2011. Currently, he is Assistant Professor in the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Pakistan. His research area include Reverse Engineering, Computer Vision, Image Processing, Deep Learning and Natural Language Processing.

KHALIL ULLAH Graduated in Computer Systems Engineering from University of Engineering and Technology Peshawar Pakistan (2002–2006). He received his Master of Science (MS) in Electronics and Communications Engineering from Myongji University South Korea in 2009. In 2016, he did PhD in Biomedical Engineering at LISiN Politecnico di Torino under Erasmus Mundus Expert II fellowship. Currently acting as Assistant Professor and Head of Software Engineering Department University of Malakand. His research is focused on extracting muscle anatomical and physiological information from High-Density Electromyography, Computer Vision, Digital signal and Image Processing, and Deep Learning with applications to medical healthcare.

ABDALLAH NAMOUN (Member, IEEE) received a bachelor's degree in computer science and a PhD in informatics from the University of Manchester, UK., in 2004 and 2009, respectively. He is an Associate Professor of intelligent interactive systems and the Head of the Information Systems Department, Faculty of Computer and Information Systems, Islamic University of Madinah. He has authored more than 50 publications in research areas spanning intelligent systems, human-computer interaction, software engineering, and technology acceptance and adoption. He has extensive experience in leading complex research projects (worth more than 21 million Euros) with several distinguished SMEs, such as SAP, BT, and ATOS. He has investigated user needs and interaction with modern interactive technologies, the design of composite software services, and methods for testing the usability and acceptance of human interfaces. His research interests include integrating state-of-the-art artificial intelligence approaches in designing and developing interactive systems.

AHMAD KHAN received a B.Sc. degree in Computer Systems Engineering from the University of Engineering and Technology (UET) Peshawar, Pakistan, in 2006, and M.Sc. degree in Computer Software Engineering from the University of Engineering and Technology, Peshawar Pakistan, in 2014. He completed PhD degree in computer systems engineering from UET, Peshawar Pakistan in October 2020. He is currently an Assistant Professor in Software Engineering, at Mirpur University of Science and Technology Mirpur AJK Pakistan. His research interests Machine to Machine Communication, the Internet of Things(IoT), and computer vision.

IBRAR HUSSAIN did his BCS (Hons) from the Department of Computer Science, University of Peshawar, Pakistan and MS (Computer Science) from the Department of Computer Science, Shaheed Benazir Bhutto University, Sheringal, Dir Upper. He completed Ph.D degree from the Department of Computer Science & Information Technology, University of Malakand, Chakdara, Dir Lower, Khyber Pakhtunkhwa, Pakistan.