Journal of Theoretical and Applied Information Technology
31st July 2022. Vol.100. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

DEEP LEARNING BASED MODEL FOR TABLE DETECTION CONTENT AND LAYOUT ANALYSIS IN COMPRESSED DOCUMENT IMAGES - A COMPREHENSIVE APPROACH

[1] KAVITA V. HORADI, [2] DR. JAGADEESH PUJARI, [3] NARASIMHA PRASAD BHAT

[1]Assistant Professor, Department of ISE, Global College of Engineering and Technology, Bangalore, India

[2] Professor, Department of ISE, SDM College of Engineering and Technology, Dharwad, India

[3] Principal Consultant, Infosys, Brookfield WI, USA

Email: [1]kavitaresearch8@gmail.com, [2]jaggudp@gmail.com, [3]bhatnp@gmail.com

ABSTRACT

Nowadays, the digital data is generated abruptly in the form of digital documents. Generation of large volumes of digital data giving rise to the big data problems has invited various problems which research community need to address. Exponential increase in the capacity of 'Big-data' containing images, textual information, audios and video content has paved a way to many challenges in processing because of an unstructured content. Due to large number of indexing and analyzing these images becomes a challenging issue. As there are various compression techniques available worldwide, these document images may undergo any compression before storage or transmission due to space and bandwidth issues. Once a document image is compressed it generates a compressed document image (CDI) which will have complexity in processing due to the loss of vital information present in it. Moreover, recognizing the layout of these documents is an important stage for various applications thus document layout analysis and recognition is considered as a promising solution for various computer vision based applications. Currently, deep learning schemes are widely adopted and comparative analysis has proven the accuracy of deep learning schemes. However, the accuracy of these systems is affected due to unstructured form of data. To overcome this issue, we present a novel scheme for layout and content equivalence analysis in compressed domain. The proposed approach uses a deep learning technique for detecting a table and faster RCNN based model for identifying the ROI. Moreover, this model incorporates the contextual information to improve the detection accuracy corresponding to each label in ROI. The proposed approach is tested by using publically available PubLayNet dataset. the average precision of PubLayNet dataset is obtained as 97.50%, F1-socre for DocBank is obtained as 97.09% and 96.55 mAP for DocBank. The comparative analysis proves that the proposed novel method attains better performance when compared with existing schemes.

Keywords: *Deep Learning; Compressed Document Images (CDI); Table Detection; Layout Analysis; Content Analysis; Faster RCNN, Datasets.*

1. INTRODUCTION

Generally, the document images consists of several semantic units or regions mainly lines, words, titles, text blocks, paragraphs, figures and tables etc. These units can be assigned with some labels to discriminate for specific tasks. Currently, the document processing for document structure and analysis is a hot research topic due to its diverse advantages in various real-time applications. This process of structure and layout analysis helps to decompose the document image into several components to understand the functional role and relationships of its various components [1].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

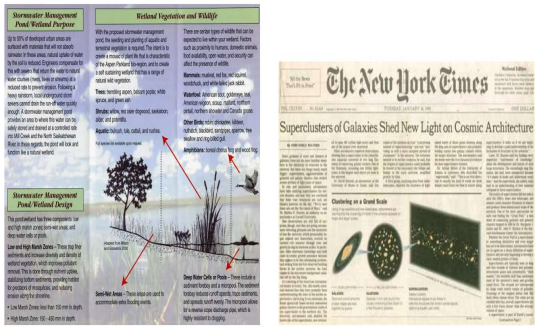

Figure.1. Example of Complex Document Layout

The document images are captured by digitizing the physical document images with the help of electronic devices, scanners and digital cameras. General document images used in day-to-day lives are usually compressed namely newspapers, brochures, magazines, invoices etc. These document images are composed of complex layouts where figures, titles, captions are placed at complex positions and these documents contains complex backgrounds and artistic text formatting as depicted in figure 1. Automated layout analysis is a promising technique [2]. The physical layout of document provides the information of physical location and boundaries of various entities in document images [4]. The manual investigation of document layout requires more time and resources to accomplish the analysis tasks. Moreover, the human processing systems are prone to errors. Thus, these systems are not suitable to process huge amount of data. As mentioned before, the document layout processing techniques decompose the image in various entities such as mainly text lines, words, characters, titles, text blocks, paragraphs, background, figures and tables. Based on the working principle, the layout processing techniques can be classified into three main categories namely bottom-up approach, top-down approach and hybrid approach for layout analysis.

- Bottom-up approach: in bottom-up approach, the algorithm considers the smallest component of document and groups these components to generate the larger, homogeneous regions. These smallest components can be in the form of pixels or connected components [5, 6].
- Top-down approach: according to this scheme, the algorithms start with complete document image and decompose the image into smaller region, iteratively [6, 7].
- Hybrid approach: in this process, both bottom-up and top-down techniques are considered for processing the document images.

1.1. NEED OF RESEARCH

Compressed domain processing involves converting a spatial domain into a frequency domain counterpart. This is done by converting the spatial domain process and the compression transforms into linear operators and forming the composite operator. The augmented digital data demand for huge storage space and adequate amount of processing capacity. Moreover, in various applications several tasks are performed on these data multiple times which becomes a resource consuming process. On the contrary, the compressed domain image processing facilitates these computations with less resource requirement. Moreover, this type of compressed domain processing can be helpful for medical and other computer vision based applications. Thus, it becomes an important aspect for computer research field.

The document images contain several entities as mentioned before which contains notion of reading order to manage the sequence of textual contents to ease the comprehension of document content. Moreover, some languages such as Chinese, Arabic etc. [2]. follows the different reading direction which creates complexities for layout analysis. The entities and their inter-relationships are known as the logical-structure of the document images which is useful to perform the segmentation. However, the variations in document text, background, layout and complexities make difficulty to achieve the appropriate outcome of layout analysis by using existing schemes. In order to overcome these issues, deep learning based schemes are widely adopted and deep learning schemes have reported significant improvement in the performance of layout analysis. In [8] Oliveira et al discussed a deep learning based technique for document image segmentation. This approach mainly focusses on page extraction, base line extraction, layout analysis and photograph extraction. This technique is based on the pixel-

wise predictor using CNN followed by task dependent post processing.

1.2. PROBLEM STATEMENT

The conventional algorithms of image processing are not suitable due to storage and resource consumption issues. The new era adopted various compressed domain processing based algorithms but these techniques did not address the content matching related issues which can be used for image indexing purpose. Moreover, the conventional schemes don't provide satisfactory performance for image matching whereas currently, deep learning based schemes has proven their significance in computer vision applications.

In this work, we focus on aforementioned issues and presented a combined scheme which considers compressed domain processing to minimize the storage requirement and other resources and RCNN based deep learning scheme for layout matching.

Rest of the paper is organized as follows: Section II highlights the literature review related to the existing techniques on document images, section III discusses on the proposed deep learning based approach to deal with the existing issues in document image segmentation, section IV summarizes the experimental analysis using the proposed methods and also presents the comparative analysis to highlight the robustness and the performance of proposed approach, and finally, section V determines the concluding notes followed by future directions.

2. LITERATURE SURVEY

This part of the paper provides the brief summary about existing techniques of finding the equivalence between document layouts. In previous section, we have described that the existing techniques suffer from the poor accuracy of matching the equivalency of document content matching where the issues of existing schemes can be resolved by applying deep learning based schemes. Moreover, document equivalence detection and analysis has several applications such as information retrieval, digital evaluation, plagiarism detection, document indexing, machine translation, optical character recognition, structural data information extraction from document images and so on. Due to its diverse applications, the document layout and content analysis plays a vital role in processing a document image. However, the existing schemes suffer from various challenging issues such as block locations, inter-and intra-class variability, and background noises.

In [9] authors described that the automated the equivalence analysis of document images requires specific-domain knowledge based information, understanding of graphs & images in the document image, and table extraction. Moreover, the current techniques fail to accurately detecting the bounding boxes and classifying them into their corresponding class. In this work, authors focused on CNN based deep learning process and presented a CNN based architecture for document layout analysis by considering tables, figures, text. Moreover, this approach uses pattern of table and text blocks and organize them in 1D signatures to minimize the dimensions of analysis. Mainly, this scheme divides the document image into multiple segment blocks and observes the tiles from each block. On these tiles, 1D project are computed to train the CNN based model and later layout classes are predicted and voting system is applied to obtain the layout of the document image.

In [10] Wu et al. reported that existing schemes of document layout analysis are directly applied on the color channels to obtain the information however, these techniques do not focus on processing the high-frequency structures in document images especially on the edge information. In order to deal with this issue, authors developed a novel approach which deals with edge information to improve the performance of document layout examination with the help of Explicit Edge Embedding Network (E3 Net). This module contains edge embedding block and dynamic skip connection block to generate the detailed features. The edge embedding network helps to include the edge information obtained from the document image. Similarly, the dynamic skip

connection blocks focus on learning the color and edge information with the help of learning weights.

Palm et al. [11] developed a new deep learning based model named as cloudscan for invoice layout analysis. Instead of relying on the set of invoice layouts, it learns the single global model of invoices which can be used as generalized layouts, thus, it can handle the unseen layouts. This uses the feedbacks obtained from various users for training purpose thus it mitigates the need of manual and precise annotation. A recurrent neural network (RNN) model is described to capture the information and test it with logistic regression model.

Agombar et al. [12] reported that existing deep learning based document image segmentation schemes are not suitable for dense layouts. Due to these issues, the data gets duplicated, lost or inaccurate which affects several application scenarios such as document image reterival. To overcome with these issues, authors developed a new segmentation scheme based on clustering mechanism.

Kosaraju et al. [13] discusses a novel method for layout analysis of scanned document images. In this process, the document image is segmented into multiple blocks. In order to achieve this, authors developed a novel texture based convolution neural network (CNN) based approach for document layout analysis called DoT-Net. The DoT-Net is a multi-classifier approach which is able to identify various blocks such as textual information, table, image, mathematical expressions and diagrams of any document image whereas the existing schemes focused only on text and non-text block classification problems.

In [14] BinMakhashen et al. discussed that historical manuscripts contain huge information which can hinder over a period of time. This degradation in image quality causes challenges in document indexing, categorization, retrieval and many other applications. Based on these considerations, authors developed feature extraction based technique for historical document layout analysis. This learning free scheme is developed into two main stages such as page characterization and segmentation. The page characterization helps to locate the main-content using top-down whitespace analysis method. This analysis helps to obtain the template features which represent the writing behavior of authors. In the next stage, a moving window model is applied to obtain the manuscript page information that defines the main-content boundary precisely. In document layout analysis and content matching, identifying table and image content is a challenging technique to enhance the performance of layout analysis scenario. Based on this, Li et al. [15] introduced a new methodology called TableBank which is used for table and image detection in document images.

Riba et al. [16] presented table detection mechanism for detecting a table in invoice document images by using graph neural network. Authors have reported that existing techniques utilize raw content (recognized text) for table detection whereas this approach uses location, context and type of content which makes is purely structure perception based approach which doesn't rely on the quality of text. The graph neural network helps to identify the local repetitive structures.

3. PROPOSED MODEL

This section of the paper discusses on the proposed solution for compressed document image layout and content analysis using deep learning based approach. First of all, we focus on table and image detection and extraction using deep learning. Later, we introduce RCNN based approach for segmentation followed by Layout and content analysis.

3.1. TRANSFORMATION OF DOCUMENT IMAGES

In first phase, we focus on table detection process which is comprised of two stages. In first phase, we adopt the image transformation scheme which helps to transform the compressed domain document image as close as possible to fine-tune the RCNN models. To achieve this, a distance transform

mechanism plays an important role where it computes the distance between the white spaces and the textual content in the given compressed document image. This measurement provides a good estimation about the presence of the table. Generally, the compressed document images consist of three different color channels as R, G and B channels. In order to obtain the transforms, three different transformation measurements such as Euclidean, linear distance and Max distance transformations are applied on the document image to obtain the tree different features. Later, these channel attributes are merged to obtain the final feature vector. In this scheme, a binary document image is considered as an input followed by applying the Euclidean distance transform on blue channel, linear distance transform on green channel and max-distance transform on red channel to obtain the transformed image. Now, this transformed image is considered for further processing for table detection.

3.2. TABLE DETECTION BY USING TRANSFORMED COMPRESSED DOCUMENT IMAGES

In this phase, we process the transformed images to detect the tables and images. For this purpose, we have adopted Faster–RCNN module which is basically used for detecting objects and classifying them in natural images. The faster- RCNN model is comprised of two main modules known as region proposal network (RPN) used for predicting object boundaries and detector module. The region proposal network proposes the various regions and the obtained regions are processed through the detector module. The region proposal network is adopted from [17]. This RPN network generates the set of rectangular object proposals with their corresponding objectness score.

In this stage, a convolution feature map is considered and a small network window is designed and slided over the obtained convolutional feature maps. Further, a spatial window having size $n \times n$ of input convolution feature maps is considered by RPN. This operation decomposes the sliding window to a lower dimension features. Later, the obtained feature maps are fed into various layers such as two fully connected layers, regression and classification layer. The two fully connected layers of this network system are shared through all spatial locations. Generally, the faster RCNN is implemented with a $n \times n$ convolution layers and followed by two 1×1 convolutional layers to obtain the regression and classification efficiency. Further, the faster R-CNN model considers the sliding window and predicts the multiple region proposals for each location which is denoted by k. The convolution operation is defined as shown in Eq.1:

$$y(p_0) = \sum_{p_n \in \mathcal{R}} w\,(p_n).\,x(p_0 + p_n + \Delta p_n \qquad (1)$$

where $\mathcal{R}$ denotes the any kernel of size $n \times n$, w is the weight of the considered kernel, x denotes the map of input attributes p_0 is the initial point of each kernel, p_n is updating the position of each kernel with all position in $\mathcal{R}$, Δp_n is the offset which are added to the normal convolution operation.

Based on these attributes, we present an amalgamation of two approaches to detect cells of the table. The aforementioned method is used to detect the multiple rows and multiple columns to obtain the cell information. To achieve this objective, we consider a feature pyramid network model with a top-down pathway approach which disseminates the information related to the high-level semantics into low-level attribute maps. An end-to-end training is presented for detecting the cell along with a training loss function. During the learning process, we try to find the coordinates of bounding box of cells as shown in Eq. 2

$$\begin{aligned} L_1 &= \sum_{r \in FR} \sum_{c_i, c_j \in r} \left\| y1c_i - y1c_j \right\|_2^2 \\ L_2 &= \sum_{r \in LR} \sum_{c_i, c_j \in r} \left\| y2c_i - y2c_j \right\|_2^2 \\ L_3 &= \sum_{r \in FC} \sum_{c_i, c_j \in c} \left\| x1c_i - x1c_j \right\|_2^2 \\ L_4 &= \sum_{c \in LC} \sum_{c_i, c_j \in c} \left\| x2c_i - x2c_j \right\|_2^2 \end{aligned} \qquad (2)$$

Here, FR denotes the first row of table, LR is the last row, FC is the first column and LC is last

column of the table in the document image whereas c_i and c_j denotes the denotes the columns. Based on this information, the alignment loss to compute the table can be computed as shown in Eq.3:

$$\mathcal{L} = L_1 + L_2 + L_3 + L_4 \qquad (3)$$

By this, it can be noted that the classification layer has 2k output scores. Meanwhile, the regression layer has 4k outputs that are encoding coordinates of 'k' boxes. Further, to add on, the 'k' regional proposals are parametrized in relevance to 'k' reference boxes also called anchors. K=9 anchors are obtained at each sliding position by Faster R-CNN. Also, it is observed that Faster R-CNN produces region proposals which are scale invariant and translational invariant. The Region Proposal Network is trained end-to-end by Stochastic Gradient Descent (SGD) and back propagation (BP). In this work, we have fine-tuned all the layers by ZF network.

3.3 LAYOUT AND CONTENT EQUIVALENCE ANALYSIS IN COMPRESSED DOCUMENT IMAGES

This section presents the layout extraction, analysis and matching process using deep learning. In order to achieve this objective, we adopt the faster-RCNN model. The faster RCNN model helps to retrieve the details from compressed document images and then analyses the layout of the document with the help of additional contextual features. In general, the deep learning model detects the major regions of article images and classifies its entities according to their corresponding labels such as main body text, table, figure, abstract, title, figure captions and others. Below given figure 2, illustrates the overall architecture of layout analysis using deep learning approach.

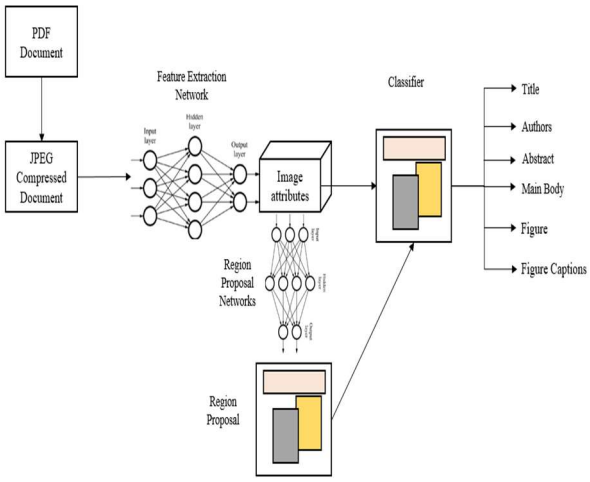

Figure.2. Overall Architecture of Layout Analysis Approach

According to this figure 2, first of all the documents are converted into JPEG images for layout analysis. The JPEG image is processed through the feature extraction network which generates the contextual features. These features are used for table and image detection and extraction with the help of region proposals. This data is processed through the classification module which generates the final outcome as labels of the detected type of entity.

In this work, we have considered Faster R-CNN to accomplish the objectives. The faster R-CNN method is a two stage detection model. This model considers document image as an input and classifies the candidate regions in to the predefined classes. The outcome of this is obtained in the form of bounding boxes for different entities. The faster-RCNN is comprised of several components such as deep convolutional neural network which uses ResNet-101 to perform the feature extraction from the input document image. This stage generates a feature map that is used to feed into region-proposal network to generate the candidate object bounding boxes. Further, these bounding boxes are fed to the classification network which assigns the predicted label to the candidates. Later, a regressor network is used which helps to fine tune the obtained bounding boxes to minimize the error in bounding boxes. The standard faster RCNN model uses attributes of image which are in the region of

interest to identify the label of that region. Generally, the Faster RCNN based object detection based techniques use the image attributes obtained from the region of interest to assign the label to that region. This module helps to improve the localization and detection accuracy because the desired content can be located at any place in the region proposals. However, in document image layout detection the detected objects are properly structured, thus incorporating contextual information to determine the label of detected region which is an important and useful task. In order to detect this, we incorporate contextual information about pages and ROI bounding boxes. These boxes are considered as important region of this process and the obtained bounding boxes are encoded as additional attributes for classification and fine tuning the regression stages. After feature extraction phase, we encode the attributes of proposed bounding box as $b = (c, w, h)$ whose centre point is $c = (c_x, c_y)$. the keypoint regression target t can be expressed as shown in Eq.4:

$$t = (t_x, t_y) = \left(\frac{k_x - c_x}{w}, \frac{k_y - c_y}{h}\right) \quad (4)$$

where $k = (k_x, k_y)$ denotes the ground truth of keypoints which are assigned to a bounding box, t_x and t_y are always in the range $[-1,1]$.

In document images, the size and number of candidates in the region of interest (ROI) contain the valuable contextual information corresponding to the labels of that region. The contextual information helps to disambiguate the region classes which are visually similar to the other regions such as abstract, body text, and figure boxes are similar to each other without context. Sometimes, more number of regions overlap thus based on their size and positon one can determine which is more likely to fit in the true document region more closely. Currently, YOLOv3 and RetinaNet are the newly developed techniques for detecting the objects and performing segmentation which are based on the single-stage detection paradigm. In this work, a Faster R-CNN model is adopted as base line model which uses separate region proposal networks and classification model which helps to capture the ROI dimensions and embed these attributes as an additional feature for classification and regression of bounding boxes.

4. RESULTS AND DISCUSSION

This section highlights the experimental analysis of the proposed approach and compares the obtained performance with the several state-of-the-art algorithms. The proposed approach is implemented using Python running on windows platform.

The performance of the proposed approach is tested on different types of document layout analysis datasets. These benchmark datasets are Article Regions [18], PubLayNet [19], and DocBank [20]. A brief discussion about these dataset is mentioned below:

PubLayNet [19]: Publaynet is a huge dataset relased by IBM. This repository contains over 1 million PDF documents which are collected from PubMed Central. In this dataset, there are over 360k page samples which includes typical document layout such as title, figure, textual data, table and list. Table.1 depicts different types of document layout categories from PubLayNet dataset.

Table.1. Document Layout Categories from PubLayNet Dataset

Category of Document Layout	XML category
Textual Data	Authors, affilitation, article informaiton,copyright description, abstract, paragraphin main text, foot note, figure and table captions
Title	Manuscript title, standalone, subsection title
List	List
Table	Table Body
Figure	Main body of figure

Table.2. DocBank Dataset Description

Split	Abstract	Author	Caption	Equation	Figure	Footer	List	Paragraph	References	Section	Table	Title
Train	25387	25909	106.723	161,140	90,249	38,482	44,927	398,086	44,813	180,774	19,638	21,688
	6.35%	6.48	26.68%	40.29%	22.61%	9.62%	11.23%	99.52	11.20%	45.19%	4.91%	5.42%
Validation	3,164	3,286	13,443	20,154	11,463	4,804	5609	49759	5549	22666	2374	2708
	6.33%	6.57%	26.89%	40.31%	22.93%	9.61%	11.22%	99.52%	11.10%	45.33%	4.75%	5.42%
Test	3176	3277	13476	20244	11378	4876	5553	49762	5641	22384	2505	2729
	6.35%	6.55%	26.95%	40.49%	22.76%	9.75%	11.11%	99.52%	11.28%	44.77%	5.01%	5.46%
All	31,727	32472	133642	201538	113270	48162	56089	497607	56003	225824	24517	27125
	6.35%	6.49%	26.73%	40.31%	22.65%	9.63%	11.22%	99.52%	11.20%	45.16%	4.90%	5.43%

DocBank [20]: The DocBank dataset is a publically available dataset containing 500K document pages having 12 different types of semantic units. These semantic units are Caption, Abstract, Equation, Title, Author, Figure, Footer, List, Paragraph, Reference, Section and Table. The table 2 describes the percentage of pages in each semantic unit.

Similarly, table 3 provides the complete distribution of document images across various years. This table provides the yearly information of number of papers for each year.

Further, we measure the performance of proposed approach for these dataset and compared the obtained performance with state-of-art techniques. We use precision, recall and f1-score evaluation metrics to measure the performance of the proposed approach. These metrics can be computed using Eq.5-Eq.7.

Table.3. Yearly Data Distribution

Year	Train		Validation		Test		All	
2014	65976	16.49%	8270	16.54%	8112	16.22%	82358	16.47%
2015	77879	19.47%	9617	19.23%	9700	19.40%	97196	19.44%
2016	87006	21.75%	10970	21.94%	10990	21.98%	108966	21.79%
2017	91583	22.90%	11623	23.25%	11464	22.93	114670	22.93%
2018	77556	19.39%	9520	19.04%	9734	19.47%	96180	19.36%
Total	400000	100.00	50000	100%	50000	100%	500000	100%

$$Precision = \frac{Area\ of\ Ground\ truth\ tokens\ in\ detected\ tokens}{Area\ of\ all\ detected\ tokens} \quad (5)$$

$$Recall = \frac{Area\ of\ Ground\ truth\ tokens\ in\ detected\ tokens}{Area\ of\ all\ ground\ truth\ tokens} \quad (6)$$

$$F1\ score = \frac{2 \times precision \times recall}{Precision + Recall} \quad (7)$$

Further, table 4 depicts the comparative analysis study in terms of F1-score and average of obtained F-score values for DocBank Dataset. The F1-score value is evaluated for each semantic unit presented in Docbank dataset.

Similarly, we measured the performance for article region dataset for sematic units which are title, caption, abstract, author, figure caption, table caption, body, figure, table, and reference. Table 5 shows the study of comparative analysis for this dataset where we compared the performance with the Faster-RCNN and VSR techniques.

Furthermore, we compared the performance for PubLayNet dataset for 5 types of semantic units

Table.4. Comparative Performance for DocBank Dataset

Models	Abstract	Author	Caption	Equation	Figure	Footer	List	Paragraph	References	Section	Table	Title	Avg
BERT [21]	0.9294	0.8484	0.8629	0.8152	1.00	0.7805	0.7133	0.9619	0.9310	0.9081	0.829	0.944	0.877
RoBERT[21]	0.9288	0.8618	0.8944	0.8248	1.00	0.7805	0.7353	0.9646	0.9341	0.9337	0.8389	0.9511	0.8891
LayoutLM [22]	0.9816	0.8595	0.9597	0.8947	1.00	0.8014	0.8948	0.9788	0.9338	0.9598	0.8633	0.9579	0.9315
BERT_LARGE [21]	0.9286	0.8577	0.8650	0.8177	1.00	0.8957	0.6960	0.9619	0.9284	0.9065	0.8320	0.9430	0.8765
RoBERT_LARGE [21]	0.9479	0.8724	0.9081	0.8370	1.00	0.7814	0.7451	0.9665	0.9334	0.9047	0.8494	0.9461	0.8988
LayoutLM_LARGE [22]	0.9784	0.8783	0.9556	0.8974	1.00	0.8392	0.9004	0.9790	0.9332	0.9596	0.8679	0.9552	0.9350
X101 [20]	0.9717	0.8227	0.9435	0.8938	0.8812	0.9146	0.9051	0.9682	0.8798	0.9412	0.8353	0.9158	0.9051
X101+LayoutLM[20]	0.9815	0.8907	0.9669	0.9430	0.9990	0.9029	0.9300	0.9843	0.9437	0.9644	0.8818	0.9575	0.9478
X101+LayoutLM_LARGE [20]	0.9802	0.8964	0.9666	0.9440	0.9994	0.9292	0.9293	0.9844	0.9430	0.9670	0.8875	0.9531	0.9488
Proposed Approach	0.9877	0.9512	0.9617	0.9611	1.00	0.9658	0.9514	0.9915	0.9612	0.9811	0.9570	0.9817	0.9709

Table.5. Comparative Analysis Based on Different Semantic Units of DocBank Dataset

Method	**Title**	**Author**	**Abstract**	**Body**	**Figure**	**Figure caption**	**Table**	**Table caption**	**Reference**	**mAP**
Faster RCNN [18]	-	1.22	-	87.49	-	-	-		-	46.38
F-RCNN context [18]	-	10.34	-	93.58	-	-	-		-	70.3
F-RCNN reimplement [18]	100	51.1	94.8	98.9	91.8	91.8	97.3	67.1	90.8	87.3
F-RCNN context reimplement [18]	100	60.5	90.8	98.5	91.5	91.5	97.5	64.2	91.2	98.8
VSR [23]	100	94	95	99.1	94.5	94.5	96.1	84.6	92.3	94.5
Proposed Approach	100	96.85	97.55	97.50	96.20	95.10	98.5	89.52	96.30	96.55

Table.6. Average Precision for PubLayNet Dataset

Method	Dataset	Text	Title	List	Table	Figure	AP
F-RCNN	Val	91	82.6	88.3	95.4	93.7	90.2
Mask RCNN		91.6	84	88.6	96	94.9	91
VSR		96.7	93.1	94.7	97.4	96.4	95.7
F-RCNN	Test	91.3	81.2	88.5	94.3	94.5	90
Mask RCNN		91.7	82.8	88.79	94.7	95.5	90.7
VSR		96.69	92.27	94.55	97.03	97.90	95.69
Proposed Approach		98.22	95.58	96.40	98.10	98.55	97.50

which are textual data, list, title, figure and table. Table 6. shows the comparative analysis for PubLayNet dataset in terms of average precision.

The experimental analysis shows the comparative analysis for detection of layout in the document image as title, list, table and figure. The comparative study shows that the proposed approach achieves the average performance as 98.22%, 95.58%, 96.40%, 98.10%, and 98.55% for Text, Title, List, Table, and Figure respectively. For this experiment, the average precision is obtained as 97.50%.

We have presented a comparative study where the performance of proposed approach shows a significant improvement in the performance. The proposed approach is capable to handle complex layouts as well as it has fast computing process due to adoption of deep learning which also helps to improve the overall performance of the system.

5. CONCLUSION AND FUTURE DIRECTIONS

In this work, mainly we have focused on Document analysis by detecting the table present in the document image. Further, a layout and content equivalence analysis in compressed document images is presented using a deep learning based framework to address this problem. Processing in the compressed domain has several advantages over storage and transmission as the decompression stage is eliminated. Hence, we have incorporated deep learning scheme which provides good performance for layout matching. The proposed visual layout analysis scheme is based on the state-of-the-art techniques of object detection model and the new architecture is obtained by incorporating the new optimization strategies. In this work, we have adopted faster RCNN based deep learning image processing for object recognition. In order to use it for compressed document images, we have incorporated contextual information of the compressed document images such as size of bounding boxes of abstract, title, tables and text. This approach uses separate region proposal networks along with faster RCNN to capture the ROI and embed the obtained features for classification. As document images are digitally born with varying complexities, future work can be on large datasets comprising of complex document images involving several components with a combination of printed and handwritten text content. As there is less exposure to the compressed domain works, this work benefits the researchers to explore more options in the field of compressed domain processing and opens up further research challenges in the area of document image processing, pattern recognition and NLP with intent to process the text and non-text content efficiently in the compressed version of the documents images.

REFERENCES

[1] Binmakhashen, G. M., & Mahmoud, S. A. (2019). Document layout analysis: A comprehensive survey. ACM Computing Surveys (CSUR), 52(6), 1-36.

[2] Amer, I. M., Hamdy, S., & Mostafa, M. G. (2017, December). Deep Arabic document layout analysis. In 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS) (pp. 224-231). IEEE.

[3] Soto, C., & Yoo, S. (2019, November). Visual detection with context for document layout analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 3464-3470).

[4] Tran, T. A., Oh, K., Na, I. S., Lee, G. S., Yang, H. J., & Kim, S. H. (2017). A robust system for document layout analysis using multilevel homogeneity structure. Expert Systems With Applications, 85, 99-113.

[5] Wu, X., Hu, Z., Du, X., Yang, J., & He, L. (2021, July). Document Layout Analysis via Dynamic Residual Feature Fusion. In 2021 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

[6] Soua, M., Benchekroun, A., Kachouri, R., & Akil, M. (2017, May). Real-time text extraction based on the page layout analysis system. In Real-Time Image and Video Processing 2017 (Vol. 10223, p. 1022305). International Society for Optics and Photonics.

[7] Barakat, B. K., & El-Sana, J. (2018, March). Binarization free layout analysis for arabic historical documents using fully convolutional networks. In 2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR) (pp. 151-155). IEEE.

[8] Oliveira, S. A., Seguin, B., & Kaplan, F. (2018, August). dhSegment: A generic deep-learning approach for document segmentation. In 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR) (pp. 7-12). IEEE.

[9] Augusto Borges Oliveira, D., & Palhares Viana, M. (2017). Fast CNN-based document layout analysis. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 1173-1180).

[10] Wu, X., Zheng, Y., Ma, T., Ye, H., & He, L. (2021). Document image layout analysis via explicit edge embedding network. Information Sciences, 577, 436-448.

[11] Palm, R. B., Winther, O., & Laws, F. (2017, November). Cloudscan-a configuration-free invoice analysis system using recurrent neural networks. In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR) (Vol. 1, pp. 406-413). IEEE.

[12] Agombar, R., Luebbering, M., & Sifa, R. (2020, August). A Clustering Backed Deep Learning Approach for Document Layout Analysis. In International Cross-Domain Conference for Machine Learning and Knowledge Extraction (pp. 423-430). Springer, Cham.

[13] Kosaraju, S. C., Masum, M., Tsaku, N. Z., Patel, P., Bayramoglu, T., Modgil, G., & Kang, M. (2019, September). DoT-Net: Document layout classification using texture-based CNN. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 1029-1034). IEEE.

[14] BinMakhashen, G. M., & Mahmoud, S. A. (2020). Historical document layout analysis using anisotropic diffusion and geometric features. International Journal on Digital Libraries, 1-14.

[15] Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., & Li, Z. (2020, May). Tablebank: Table benchmark for image-based table detection and recognition. In Proceedings of the 12th Language Resources and Evaluation Conference (pp. 1918-1925).

[16] Riba, P., Dutta, A., Goldmann, L., Fornés, A., Ramos, O., & Lladós, J. (2019, September). Table detection in invoice documents by graph neural networks. In 2019 International Conference on Document Analysis and Recognition (ICDAR) (pp. 122-127). IEEE.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99

[18] Soto, C., Yoo, S.: Visual detection with context for document layout analysis. In: EMNLP-IJCNLP. pp. 3462–3468 (2019)

[19] Zhong, X., Tang, J., Jimeno-Yepes, A.: Publaynet: Largest dataset ever for document layout analysis. In: ICDAR. pp. 1015–1022 (2019)

[20] Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., Zhou, M.: Docbank: A benchmark dataset for document layout analysis. In: COLING. pp. 949–960 (2020)

[21] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-theart natural language processing. ArXiv, abs/1910.03771

[22] https://aka.ms/layoutlm

[23] Zhang, P., Li, C., Qiao, L., Cheng, Z., Pu, S., Niu, Y., & Wu, F. (2021). VSR: A Unified Framework for Document Layout Analysis combining Vision, Semantics and Relations. arXiv preprint arXiv:2105.06220.