Article

A Survey of Graphical Page Object Detection with Deep Neural Networks

Jwalin Bhatt [1,3,†], Khurram Azeem Hashmi [1,2,3†*], Muhammad Zeshan Afzal [1,2,3†] and Didier Stricker[1,2]

[1] Department of Computer Science, Technical University, 67663 Kaiserslautern, Germany; jbhatt@rhrk.uni-kl.de (J.B); khurram_azeem.hashmi@dfki.de (K.A.H); didier.stricker@dfki.de (D.S);
[2] German Research Institute for Artificial Intelligence (DFKI), 67663 Kaiserslautern, Germany; muhammad_zeshan.afzal@dfki.de;
[3] Mindgrage, Technical University, 67663 Kaiserslautern, Germany
* Correspondence: khurram_azeem.hashmi@dfki.de
† These authors contributed equally to this work.

Abstract: In any document, graphical elements like tables, figures, and formulas contain essential information. The processing and interpretation of such information require specialized algorithms. Off-the-shelf OCR components cannot process this information reliably. Therefore, an essential step in document analysis pipelines is to detect these graphical components. It leads to a high-level conceptual understanding of the documents that makes digitization of documents viable. Since the advent of deep learning, the performance of deep learning-based object detection has improved many folds. In this work, we outline and summarize the deep learning approaches for detecting graphical page objects in the document images. Therefore, we discuss the most relevant deep learning-based approaches and state-of-the-art graphical page object detection in document images. This work provides a comprehensive understanding of the current state-of-the-art and related challenges. Furthermore, we discuss leading datasets along with the quantitative evaluation. Moreover, it discusses briefly the promising directions that can be utilized for further improvements.

Keywords: Deep neural network; survey; document images; review paper; deep learning; performance evaluation; page object detection, graphical page objects; document image analysis; page segmentation

1. Introduction

The rapid increase in digitization of document images in both financial and non-financial sectors has considerably improved the accessibility of the data. To obtain reliable information from these scanned document images, options like manual capturing of data have become highly laborious and impractical. Therefore, over the last few decades, accurate information extraction has been vital research for the document analysis community [44–47].

Apart from the text, information in scanned documents is often stored in a graphical manner such as Tables, formulas, figures. These are referred to as graphical page objects in document analysis community [21]. Figure 1 illustrates the problem that involves the detection of figures, formulas, and tables in document images. While state-of-the-art optical character recognition systems [44,48] extract the textual information conveniently, they have a hard time processing information from these graphical page objects. Hence, it is essential to develop approaches that can parse the information from these page objects.

With the recent surge of deep learning-based object detection algorithms in computer vision [49–51], a considerable amount of methods are developed that have formulated the problem of detecting graphical page objects in document images as an object detection problem. Furthermore, several datasets consisting of thousands of annotated scanned document images are also published. Although the approaches leveraging these datasets have significantly improved state-of-the-art, a consolidated comparison among these approaches is missing.

Citation: Lastname, F.; Lastname, F.; Lastname, F. Title. *Preprints* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

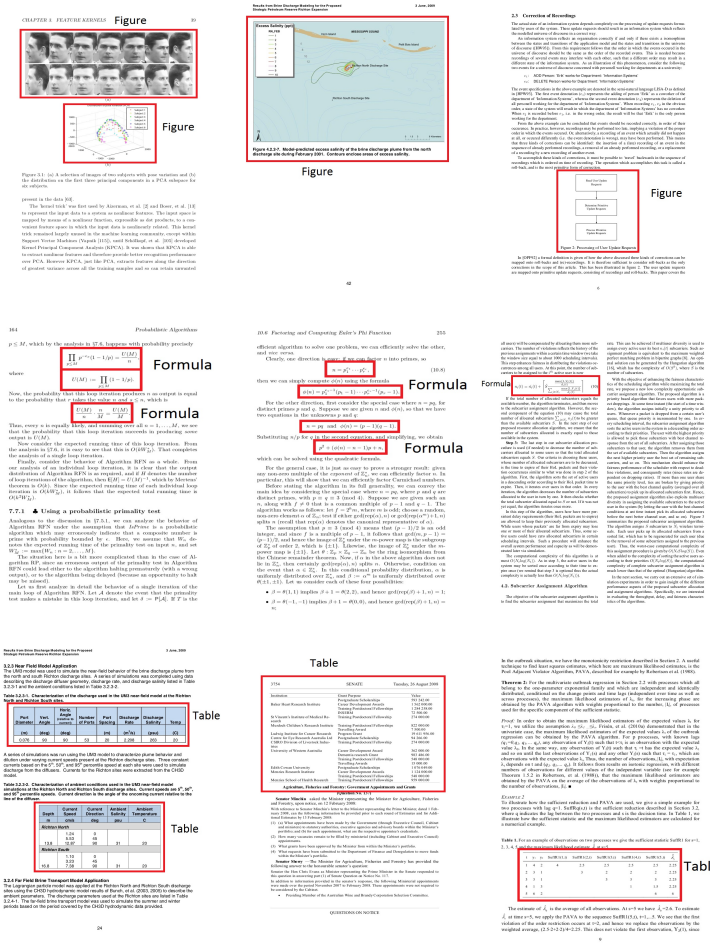

Figure 1. Demonstration of the problem of graphical page object detection in document images. **First Row**: figure detection in document images. **Second Row**: detection of single and multiple formulas in document images. **Third Row**: localization and classification of tabular areas in document images. The samples are taken from the dataset of ICDAR-17 POD [43].

In this survey paper, we have presented a thorough analysis of the recent state-of-the-art approaches that have approached the problem of graphical page object detection in scanned document images by employing deep neural networks. Since page objects can be of several types [40], we have covered the three most important page objects in document images [43]. These graphical page objects are referred to as **table**, **formulas**, and **figures**.

This paper has investigated how deep neural networks-based approaches work on detecting these types of page objects. Therefore, we have covered the most relevant approaches that have produced state-of-the-art results in this domain. Some of the discussed approaches work only on a single page object, and some have covered all three of them. However, our primary focus is to provide the perspective about the outcome of deep learning-based approaches on graphical page object detection in document images. To summarize, our contributions are as follows:

1. We present the comparisons between recently introduced algorithms for improving page object detection by highlighting their advantages and limitations.
2. We present a brief overview of the publicly available challenging datasets for graphical page object detection.
3. We provide an evaluative comparison among the state-of-the-art graphical page object detection systems.

This review paper is organized as follows: Section 2 presents a brief overview of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

prior works that have exploited traditional approaches to detect graphical page objects. Section 3 explains all the approaches contributing to graphical page objects by leveraging deep learning methods. Figure 3 depicts the organizational structure of those explained methodologies; Section 4 highlights all the publicly available datasets that can be employed to tackle the mentioned problem. Section 5 explains the mostly employed evaluation metrics and analyzes performances of all the discussed approaches in Section 3. Section 6 discusses the current challenges and presents the conclusion whereas Section 7 highlights the future directions.

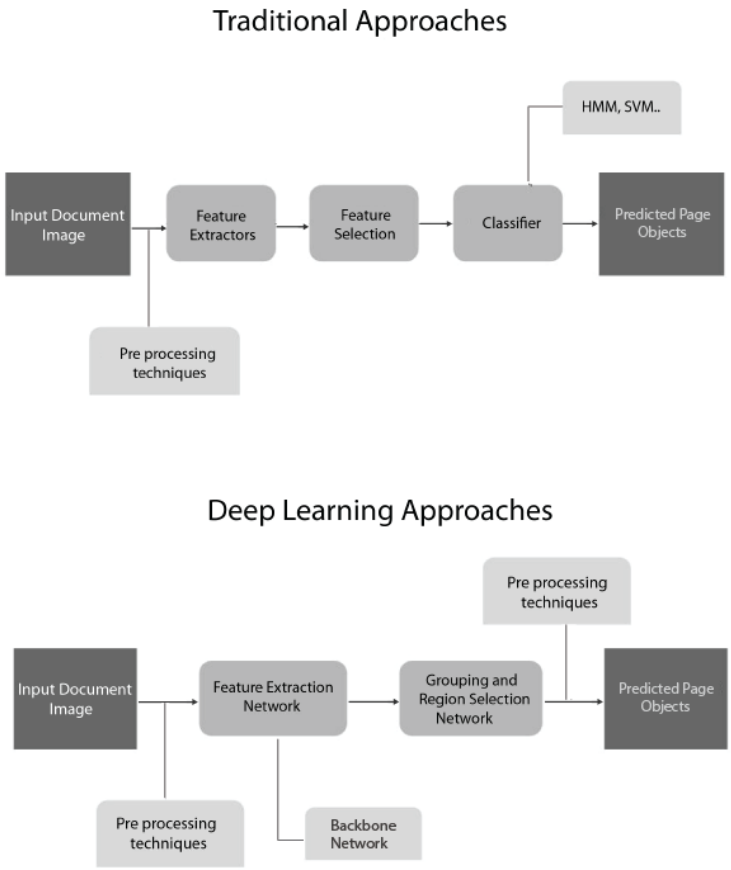

Figure 2. Visual depiction of the basic differences between the traditional methods and the deep learning-based techniques. Traditional approaches rely heavily on image processing methods and custom heuristics whereas deep learning techniques leverage convolutional neural networks-based architectures. In deep learning approaches, spatial features from the document images are extracted from backbone networks such as VGG-16 [18] or ResNet [61]. These features are further propagated to region detection or segmentation networks to classify and localize page objects.

2. Traditional Approaches

The problem of graphical page object detection in documents is a well-recognized problem. Several approaches that employed traditional methods are introduced in this domain. Figure 2 illustrates the fundamental differences between the traditional approaches and the deep learning-based approaches. The traditional approaches leverage image processing techniques such as binarization and connected component analysis. Contrarily, deep learning-based methods utilize backbone CNN to generate the spatial feature maps from the document images.

In order to implement table detection, the prior techniques [1,3,4] have defined a certain underlying structure for tables in a document. Tupaj et al. [2] employed Optical

 doi:10.20944/preprints202104.0739.v1

Character Recognition (OCR) to extract tabular information. The method tried to recognize possible table areas by analyzing the keywords and white spaces. The main disadvantage of this approach is that it is fully based on the presumptions regarding the tables' structure and the collection of the used keywords.

Wang et al. [8] proposed another approach in the field of table analysis. It utilizes distance between consecutive words to detect table lines. Subsequently, adjacent vertical lines are grouped with consecutive horizontal words to propose table entity candidates. However, the underlying assumption is that there can be a maximum of two columns in a table. Hence, three types of layouts (single, double, and mixed columns) are designed in this approach. The drawback of this method is that it is only applicable to a limited number of designed temples.

Kieninger et al. [5–7] introduced a system called T-Recs to extract tabular information from documents. Their method takes the word bounding boxes which are segregated to build a segmentation graph in a bottom-up manner. Their system is vulnerable to tables containing multi rows and columns.

A method for detecting tables by calculating the intersection area between the vertical and horizontal lines was suggested by Gatos et al. [11]. The recreation of tables is then done by denoting corresponding vertical and horizontal lines related to intersection pairs. This approach presumes that a table should have ruling lines. A method for table detection by using Hidden Markov Models (HMMs) was suggested by Costa e Silva et al. [12]. The method fetches text from PDF files by applying the pdftotext Linux utility. Then feature vectors are computed based upon the gaps present between the text. This approach can only be employed for non-raster PDF files that do not contain noisy data.

A method for table detection under the assumption that tables in documents can contain only singular columns is proposed by Hu et al. [9]. Another technique for table detection in heterogeneous documents was proposed by Shafait et al. [10]. This mechanism is built into an open-source Tesseract OCR engine [52]. Although these traditional approaches were effective on the documents with restricted layout variations, either they rely on the meta-data or highly depends on the post-processing methods involving custom heuristics. Furthermore, they fail to produce similar results on generic datasets. Therefore, it is essential to exploit recently proposed deep learning techniques to tackle the problem of graphical page object detection in document images.

3. METHODOLOGIES

Graphical objects like tables, figures, and formulas are an integral part of documents because they hold a significant amount of information in a confined space. As explained in Section 1, detecting the graphical object means localizing these objects within a document image. Conceptually this problem is identical to localizing the objects in natural scene images. Recently, deep learning algorithms have also attracted the interest of researchers in the document image analysis community.

This section will discuss the methodologies that have utilized the capabilities of deep neural networks to solve the problem of graphical page object detection in document images. By following the convention of [21], we have covered approaches that have worked on the detection of the following graphical page objects in document images: 1) Tables, 2) Figures, and 3) Formulas.

For the convenience of the readers, we have classified the methodologies according to the employed deep learning concepts. We discuss the organizational flow of the methodologies in Figure 3. Table 1 summarizes the presented approaches and highlights their advantages and limitations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

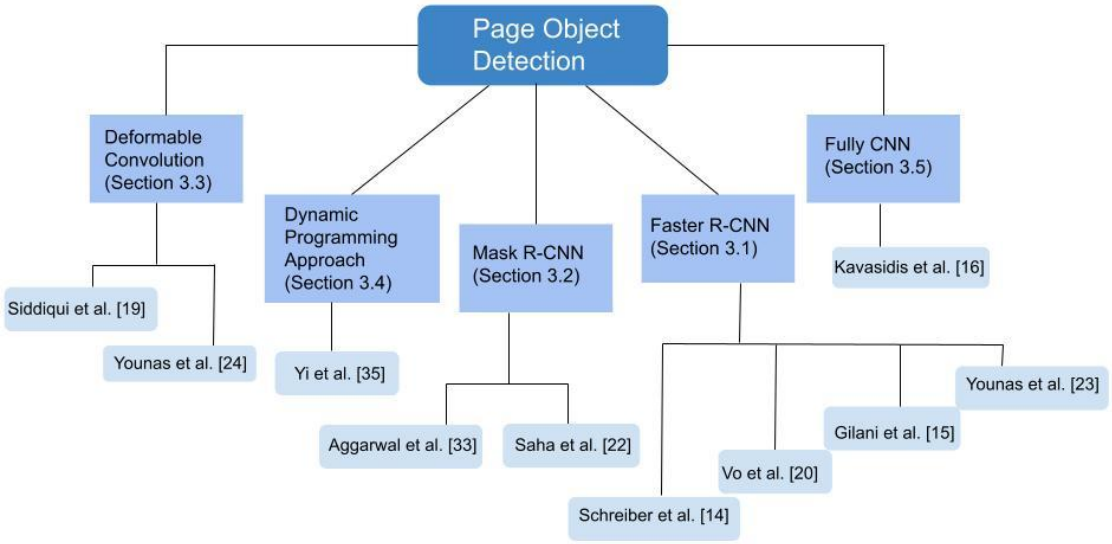

Figure 3. Categorization of discussed methodologies in the paper. The problem of page object detection is tackled through employing various deep learning concepts. The explained approaches are divided conceptually.

3.1. Faster R-CNN

Recently, it has been the case that the improvement of object detection algorithms in the field of computer vision has a direct relation with the improvement of graphical page object detection in document images. Faster R-CNN [50] which is the improved version of Fast R-CNN [53] is a two-stage object detection network. Figure 4 illustrates the architecture of Faster R-CNN. In order to obtain a detailed explanation about the architecture, readers may refer to [50]. This section covers the approaches that detect the graphical page objects by exploiting the capabilities of Faster R-CNN [50].

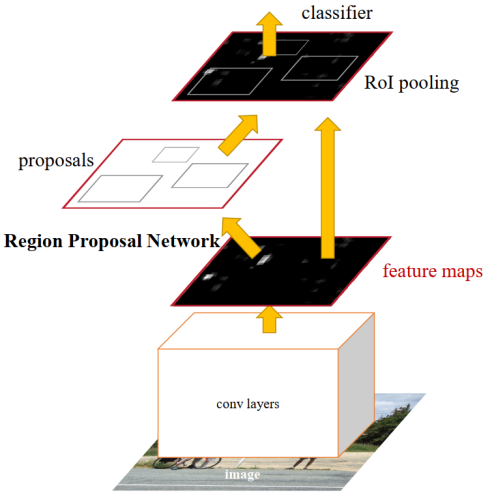

Figure 4. Explained architecture of Faster R-CNN. Image is obtained from [50]

An image-based deep learning table detection approach was suggested by Schreiber et al. [14] where they implemented Faster R-CNN for detection of tables in document images. The paper presents that the recently introduced object detectors dependent on Convolutional Neural Networks (CNN) can detect tables in document images. By leveraging back-bones like ZFNet [17] and VGG-16 [18], the authors have achieved promising results on ICDAR-13 dataset [41]. Their approach has also utilized the transfer learning technique

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

by using the pre-trained model on the Pascal-VOC dataset [54]. They also attempt table structure recognition along with table detection.

Vo et al. [20] published a method for page object detection, which involves detecting figures, formulas, and tables. Their technique makes use of an ensemble technique of fast R-CNN [53], and faster R-CNN [50]. They combined the region proposals obtained from Fast R-CNN and Faster R-CNN and then apply bounding box regression to boost performance. They have used the ICDAR-17 POD [21] dataset to benchmark their approach.

The blend of traditional methods and deep learning networks is presented by Younas et al. [23] to solve the problem of formula and figure detection in document images. The authors propose that instead of giving raw input images to object detection algorithms, transformed image representations yield better results. Connected component analysis (CC), distance transform, and color transform on the raw input images are performed and are subsequently processed using the Faster R-CNN model.

Gilani et al. [15] have utilized a similar technique. They have used the image transformation method in which a Euclidean distance transform [55], linear distance transform [56], and max distance transform [57] are applied on blue, green, and red channels of the input image, respectively. This transformed image is further propagated to Faster R-CNN to identify and regress the tabular boundaries in document images.

Another approach in which performance of two state-of-the-art object detection networks: Faster R-CNN [50] and Mask R-CNN [49] is compared on graphical page objects [22]. The article presents exhaustive evaluations on the detection of tables, formulas, and figures in document images. The paper's conclusion states that Mask R-CNN [49] is better suited to solve the problem of page object detection because of having extra components in the loss function.

3.2. Mask R-CNN

Mask R-CNN [49] is the extended model of Faster R-CNN [50] with an addition of an extra loss known as segmentation loss. Figure 5 depicts the basic architecture of Mask R-CNN. However, the comprehensive detail about the network can be found at [49]. The graphical page objects present in the document images have very low inter-class variance. An object originally labeled as a table can easily be misinterpreted with a figure or formula. By leveraging the segmentation loss of Mask R-CNN, researchers in the document image analysis community have improved the performance of graphical page object detection systems. This section covers those methodologies.

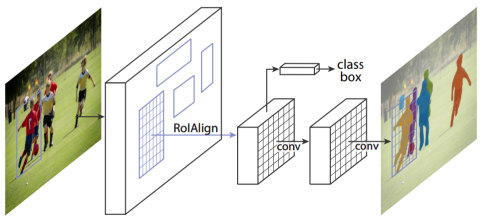

Figure 5. Explained architecture of Mask R-CNN. Image is obtained from [49]

Saha et al. [22] published the method for page object detection in document images through employing Mask R-CNN. Their end-to-end deep learning-based system, called Graphical Object Detection (GOD), detects the tables, figures, and formulas directly from the raw input images. The authors propose that there is no need to add extra pre or post-processing steps to solve page object detection. By leveraging the power of transfer learning, the authors have done bench-marking on the well-known datasets of ICDAR-17 POD [21], UNLV [58], and ICDAR-13 [41].

A recent end-to-end table detection network called CDeC-Net is introduced by Agarwal et al. [30]. The system CDeC-Net leverages the novel object detection network Cascade

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

Mask R-CNN based on Cascade R-CNN [51]. The presented article has shown a noticeable improvement in the performance of table detection system across several datasets such as ICDAR-17 POD [21], ICDAR-13 [41], ICDAR-2019, Marmot [34], TableBank [35], PubLayNet[43], and UNLV [58]. After extensive evaluations, the authors have concluded that the network Cascade Mask R-CNN is superior to the previous state-of-the-art table detection systems.

3.3. Deformable Convolutions

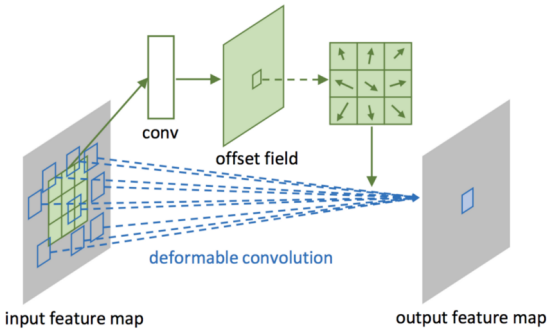

Figure 6. Architecture of 3 × 3 Deformable Convolution. Image is obtained from [36]

Deformable convolutions differentiate the conventional convolutions by providing the leverage of deformable modules. The deformable module learns the sampling matrix with the location offsets. The offsets are learned according to the previous feature maps through additional convolution layers. This process makes the receptive field dynamic and enables the convolutional filters to adapt to different scales. While Figure 6 depicts the basic intuition behind the deformable convolutional networks, thorough information about the architecture is explained in [36]. Most of the mentioned methodologies have employed conventional convolutions in their object detection frameworks to solve page object detection in document images. Recently, instead of conventional convolutions, deformable convolutions [36] are investigated to detect tables, figures, and formulas. This section highlights those approaches.

Siddiqui et al.[19] proposed an approach to detect tables that leverages deformable convolutions in their object detection framework. The authors argue that deformable convolutions are better suited for the problem of table detection. Because of their dynamic receptive field, tabular areas belonging to various scales and aspect ratios can be localized conveniently. The authors employed Faster R-CNN [50] by replacing a conventional Feature Pyramid Network (FPN) with a deformable FPN module. After extensive evaluations, the authors proved that deformable Faster R-CNN had outsmarted the conventional Faster R-CNN for the problem of table detection in document images.

Younas et al. [24] exploited a similar approach by employing a deformable FPN module to detect formulas and figures in document images. Instead of providing raw input images to their deformable Faster R-CNN model, the authors have proposed an image transformation method identical to [23]. With the combination of transformed image representation and deformable object detection architecture, the authors have produced state-of-the-art results for the figure and formula detection on the famous ICDAR-17 POD dataset [21]. Along with the novel approach, the writers have also corrected the ICDAR-17 POD dataset [21] and have made it publicly available [1].

[1] https://bit.ly/2AUSlzI

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

3.4. Dynamic Programming Based Approach

Yi et al. [32] introduced a deep learning-based graphical page object detection approach similar to the object detection algorithms. In the presented approach, a convolutional neural network designed specifically for page object detection proposed candidate regions that are refined through a dynamic programming approach instead of the well-known non-maximum suppression method [59]. Tables, figures, formulas, and text lines are localized in document images by their system. The authors argue that page objects have a high variance in their aspect ratios, unlike objects in natural scenic images. Therefore, non-maximum suppression is not well-suited to detect all the page objects in a document image. The presented work compares the performance of their system with the conventional object detection approach of Fast R-CNN [53] and Faster R-CNN [50], and concludes that the dynamic programming-based approach has outperformed the rest of the methods.

3.5. Fully Convolutional Neural Networks

Along with object detection algorithms, Fully Convolutional Neural Networks (FCNNs) [60] have been exploited to solve graphical page object detection in document images. The basic intuition behind FCNNs is assigning the label for each pixel present in an image. Figure 7 depicts the architecture of FCNNs and for further explanation, we refer our readers to [60]. Kavasidis et al. [16] posed the problem of tables and chart detection as a saliency detection problem. The authors propose that each class of page object can be referred to as a separate saliency category. To segment those categories (tables and charts), the system employs FCNNs where each pixel will be classified into tables, charts, or a background in a document image. The obtained saliency map is further propagated to the fully connected Conditional Random Field (CRF) [60], which smooths the system's output.

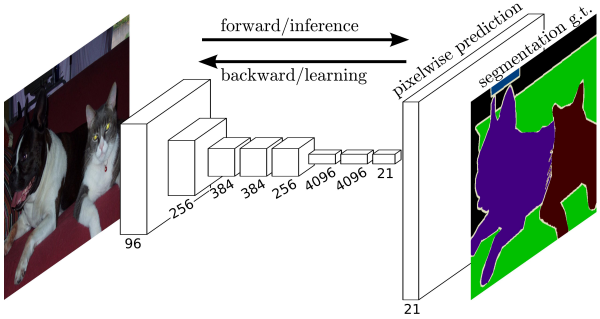

Figure 7. Explained architecture of Fully Convolutional Neural Network. Image is obtained from [63]

4. Datasets

Deep neural networks consist of a huge number of parameters. To achieve convergence, datasets with a massive amount of images are required to train these networks optimally [14,19]. Recently, the document image analysis community published several public datasets. Some of these datasets have provided annotations for various graphical page objects. This section will mainly cover the recently published datasets that contain information about the boundaries of tables, formulas, and figures. Figure 8 depicts few samples of these datasets.

Moreover, we discuss few datasets that only contain annotations for one of the three mentioned page objects, such as tables. Figure 9 depicts a couple of samples belonging to these datasets. Table 2 presents the summary of all the datasets covered in this section.

4.1. ICDAR-17 POD

ICDAR-17 Page Object Detection (POD) [21] is a publicly available dataset introduced in the page object detection competition at ICDAR 2017. This dataset is one of the most widely used datasets to evaluate graphical page object detection systems. The dataset comprises a page consisting of various layouts such as single-column, double-columns,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

Table 1: A Summary of different graphical page object detection methods that have employed deep neural networks.

Literature	Method	Highlight	Limitation
De-CNT [19]	Deformable convolutions, implemented in the Faster R-CNN [50] architecture	The dynamic receptive field of deformable CNN helps in recognizing various tabular broundaries	Deformable CNN requires more computation as compared to conventional CNN
Fi-Fo Detector [24]	Color transform, connected component analysis, distance transform applied on images that are fed to deformable pyramid network.	**a)** Transformed images yield better results as compared to raw input images. **b)** The approach leverages the deformable FPN model in their object detection network.	The approach depends on the extra pre-processing steps
DeepDeSRT [14]	Faster R-CNN [50] with transfer learning	Straightforward and effective approach to detect tables.	Does not perform as accurate as other state-of-the-art methods.
Vo et. al [20]	Ensembling of Fast R-CNN [53] and Faster R-CNN [50]	Leveraging the power of both selective search and region proposal network to generate accurate region of interests.	Computationally expensive because of combination of two separate object detection networks.
GOD [22]	Faster R-CNN [50] and Mask R-CNN [49]	Simple end-to-end approach to detect multiple page objects	The network often confuse the similar looking page object belonging to different classes.
Gilani et al. [15]	Transformed images are passed through Faster R-CNN [50]	The distance transform method helps the object detection network to focus on desired page objects	Requires extra pre processing method.
CDEC-Net [30]	Cascaded Mask R-CNN [51]	**a)** Multi-scale feature pyramid network. **b)** Deformable Convolution improves the performance.	Method requires high computational resources
Kavasidis et al. [16]	Semantic image segmentation with saliency detection.	**a)** Table detection formulated as a task of saliency detection. **b)** Dilated convolutions instead of traditional convolution improve efficiency.	The approach depends on multiple pre-processing steps to achieve good results
Yi et al. [32]	Dynamic programming based technique	Replaces non maximal suppression with dynamic programming algorithm improves the refining process for region of interests.	Extra post-processing over head.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

and multi-columns. This dataset has an annotation for tables, formulas, figures present in document images. The page objects contain headings, textual, page title, content, captions, etc. This dataset contains 2417 English document images in total, extracted from 1500 scientific papers of CiteSeer. This dataset is split into training and test set, having 1600 and 817 document images, respectively. Contents are as follows: Training set contains 1978 figures, 698 tables, and 3515 formulas, while Test set contains 961 figures, 371 tables, and 1192 formulas.

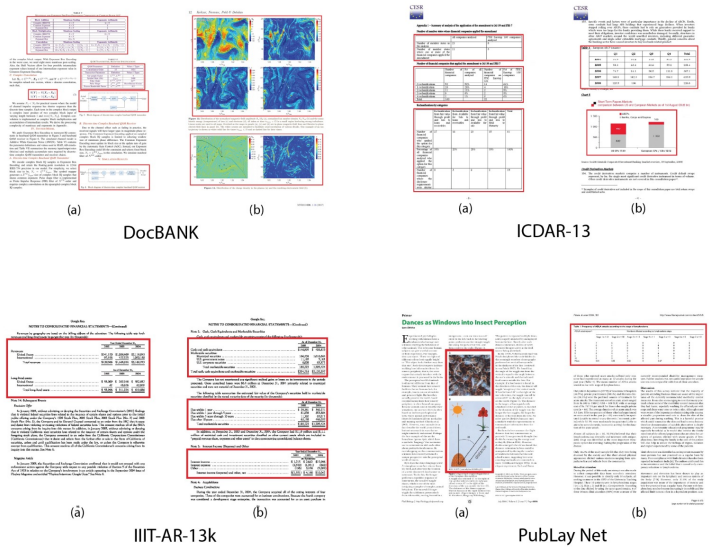

Figure 8. Sample document images taken from the various datasets of DocBank [40], ICDAR-13 [41], IIT-AR-13K [38], and PubLayNet [43]. Part (a) and (b) represent the highlighted graphical page objects in a document image.

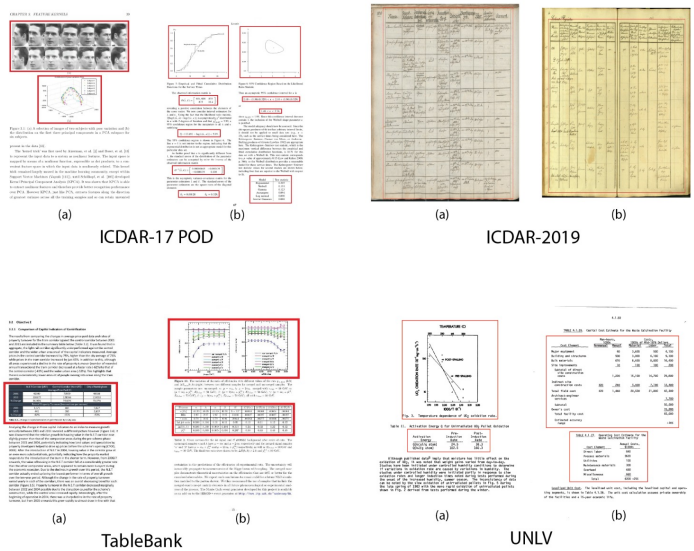

Figure 9. Sample document images taken from the various datasets of ICDAR-17 POD [21], ICDAR-19 [33], TableBank [35], and UNLV [42]. Part (a) and (b) represent the highlighted graphical page objects in a document image. It is important to mention that most of the datasets illustrated in this figure have annotations for the tabular boundaries only.

 doi:10.20944/preprints202104.0739.v1

4.2. PubLayNet

In 2019, Zhong et al. [43] published a huge dataset for document layout analysis known as PubLayNet. This dataset is generated by automatically annotating the document layout of over 1 million PubMed Central^{TM} PDF articles. The dataset contains various document layout categories, including text, title, list, table, and figures. Having more than 360 thousand annotated document images, this huge dataset facilitates the researchers to develop and evaluate advanced deep learning-based models for document page object detection.

4.3. DocBank

Another dataset to solve document layout analysis is released by Li et al. [40]. The dataset is known as DocBank, which is the extended version of the TableBank dataset [35]. DocBank is a novel large-scale dataset, and it is constructed by employing weak supervision from the LaTeX documents available on arXiv.com. The proposed dataset comprises 500 thousand document pages with 12 different kinds of semantic blocks such as tables, figures, equations, figures, lists, paragraphs, etc. The authors also define the training/val/test splits in which 400 thousand samples are used for the training purpose, whereas 50 thousand samples are allocated for validation and testing purposes. This large-scale rich dataset extends the opportunities to investigate the blend of deep neural networks employed in computer vision with the methods mainly used in document analysis.

4.4. Marmot

Marmot is widely utilized by scientists in the area of understanding the tables and formulas. This dataset has been published by the Institute of Computer Science and Technology (Peking University) and described in the paper proposed by Fang et al. [34]. This dataset consists of 2000 document images. These images are comprised of conference papers of both English and Chinese languages from 1970 to 2011. There is roughly a 1:1 ratio for both positive and negative images in the dataset. Due to the complex page layouts, this dataset is highly applicable for evaluating table detection systems. There were few instances of incorrect annotations in the dataset, which is corrected by Schreiber et al. [14]. The size of the dataset reduces to 1967 document images after the correction.

4.5. TableBank

During early 2019, in the community dedicated to table detection, Minghao et al. [35] recognize the requirement for enormous datasets and established TableBank. TableBank is a dataset consisting of 417 thousand labeled images utilizing tabular data. The dataset has been accumulated by gathering the information over the documents which are present in .docx format. The dataset also contains another form of information, that is, Latex documents which were accumulated from the arXiv 5 database. The publishers of this dataset suggest the usage of this dataset for both structural recognition and table detection tasks. The authors of this dataset claim that this large-scale dataset will enable the researchers to exploit the capacity of deep neural networks.

4.6. IIIT-AR-13k

Mondal et al. [38] have proposed a novel IIIT-AR-13k dataset. The dataset mainly consists of business-type documents. There are in total 13 thousand pages containing graphical elements like tables, signatures, figures, and so on. The bounding boxes are marked in a non-automated manner to construct the dataset. The authors generated this dataset manually, and it is one of the biggest datasets in the domain of graphical page object detection.

4.7. DeepFigures

Based on our knowledge, DeepFigures [37] is one of the most extensive free-to-use datasets to utilize for the task of graphical page object detection. It comprises more than

 doi:10.20944/preprints202104.0739.v1

1.4 million documents along with the information of boundaries of tables and figures. The authors leverage the scientific articles found online on the databases like PubMed and arXiv to create the dataset. This large-scale dataset provides an opportunity to investigate the performance of table and figure detection systems in document images.

4.8. ICDAR-13

ICDAR-13[41] is widely utilized for the problem of table detection and table structure extraction. The dataset consists of PDF files. These PDF files are converted into images. The dataset is composed of graphs, structured tables, text as information, and charts. However, It only provides annotations for structure data for table recognition and table detection. This dataset has 67 PDFs with 150 tables in which 27 PDFs are from the EU, and 40 PDFs are from the US Government. In total, this dataset has 238 images, from which 128 images contain table information. This dataset is often used for reporting and comparing.

4.9. UNLV

For document image analysis, UNLV[42] is a very well-known dataset in this field. This dataset is composed of various documents like business letters, magazines, reports, newspapers, etc. Even though this document has almost 10,000 images, only 427 images possess a tabular region. Often, the research community uses the images that contain the tabular regions to manage numerous experiments.

4.10. ICDAR-2019

In 2019, Competition on Table Detection and Recognition(cTDaR) [33] is executed in ICDAR. The competition proposes two new datasets: historical and modern datasets. The historical datasets encompass train schedules, simple tabular prints from old books, images from hand-written accounting ledgers, and so on. In contrast, modern datasets encompass samples from forms, financial documents, and scientific papers. This dataset has become a benchmark dataset to assess the performance of state-of-the-art systems for table analysis.

5. Evaluation

This section covers the well-known evaluation metrics that have been employed by the deep learning-based approaches to assess their performance and compares the results among various state-of-the-art approaches. Moreover, we will present the comprehensive evaluative comparison between the methodologies that are explained in Section 3.

5.1. Precision

The metric precision is defined as the ratio between the correctly predicted positives samples to the total positive samples. Figure 10 depicts the definition of precision for the problem of graphical page objects in document images. Mathematically, it is described as:

$$Precision = \frac{\text{correct prediction}}{\text{total predictions}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

Where TP denotes the True Positives and FP represents False Positives.

5.2. Recall

The metrics recall evaluates the performance of the system by calculating the number of corrected predictions in the actual test set. It is calculated as follows:

$$Recall = \frac{\text{correct predictions}}{\text{Total correct annotations in ground-truth}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

Where TP denotes the True Positives and FN represents False Negatives.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

Table 2: Graphical page object datasets. It is important to mention that we have considered equation and formula as semantically equal in this table. Some of these datasets contain as many as 12 page objects [40]. For the sake of convenience, we have only included table, figure, and formula.

Dataset	Table	Figure	Formula	# Samples	Year	Location
PubLayNet [43]	✓	✓	✗	360K	2019	https://developer.ibm.com/exchanges/
DocBank [40]	✓	✓	✓	500K	2020	https://doc-analysis.github.io/docbank-page
ICDAR-17 POD [21]	✓	✓	✓	2.4K	2017	https://www.icst.pku.edu.cn/cpdp
IIIT-AR-13k [38]	✓	✓	✗	13K	2020	http://cvit.iiit.ac.in/usodi/iiitar13k.php
DeepFigures [37]	✓	✓	✗	5.5K	2018	https://s3-us-west-2.amazonaws.com/ai2-s2-research-public
ICDAR-13 [41]	✓	✗	✗	238	2013	http://www.tamirhassan.com/html/
UNLV [42]	✓	✗	✗	427	2010	http://www.iapr-tc11.org/mediawiki/index.php?
ICDAR-2019 [33]	✓	✗	✗	3.6K	2019	https://zenodo.org/record/2649217
Marmot [34]	✓	✗	✓	958	2012	https://www.icst.pku.edu.cn/cpdp/sjzy
TableBank [34]	✓	✗	✗	417K	2020	https://doc-analysis.github.io/tablebank-page

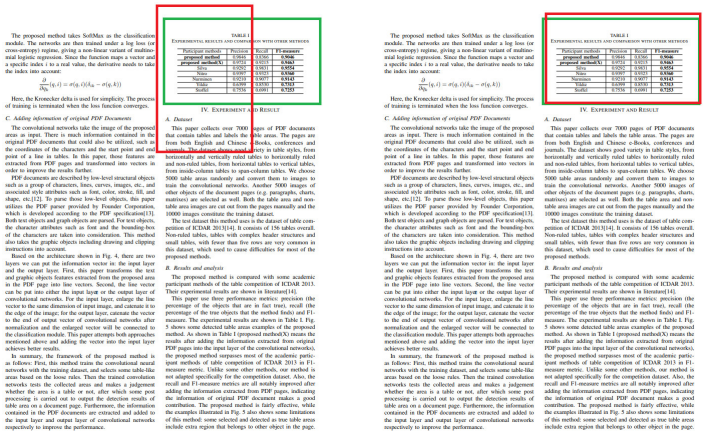

Figure 10. An instance of a precise and imprecise table detection. Green color represents the ground-truth tabular area whereas red color denotes the predicted tabular boundary.

5.3. F-Measure

The harmonic mean of precision and recall is known as F-Measure. The formula for finding an F1 score is given by:

$$F\text{-}Meausre = \frac{2\times \text{ Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{3}$$

5.4. Intersection Over Union

Intersection Over Union (IOU) is a well-known evaluation metric commonly used in evaluating the capabilities of object detection algorithms. Since object detection techniques have been widely exploited to solve graphical page object detection, we have decided to discuss this metric in our paper. IOU calculates how much the area of the predicted bounding box intersects with the area of the actual ground-truth. For the sake of convenience, an example of computing IOU is illustrated in Figure 11. Mathematically, it is explained as

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

follows:

$$IOU = \frac{\text{Area of Overlap region}}{\text{Area of Union region}} \tag{4}$$

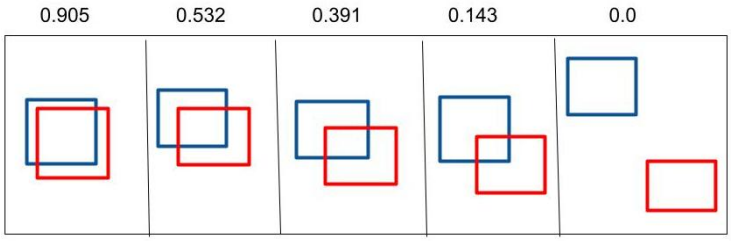

Figure 11. Visual illustration of IOU in object detection methods. The bounding box with blue color represents the ground-truth whereas the bounding box with red color denotes the predicted bounding box. Considering the IOU threshold set to 0.5, only the first two predictions from the left will be considered true positives whereas the rest of them will be treated as false positives.

The problem of graphical page object detection is to localize the boundaries of formulas, figures, and tables. For the sake of visual convenience, we have divided the performance evaluation between the explained methodologies into three separate tables. The quantitative analysis for detection of tables, figures and formulas are summarized in Table 3,4, and 5 respectively. Various approaches have evaluated their methods on distinctive IOU thresholds.

5.5. Evaluation for Table Detection

It can be observed by looking at Table 3, the instance segmentation-based architectures like Cascade Mask R-CNN has outperformed the rest of the approaches with a slight margin. It shows that the multi-scale classification module that has improved the generic object detection [62], has also advanced the table detection systems in document images.

5.6. Evaluation for Figure Detection

Table 4 compares the performance between the two recently proposed deep learning-based approaches for figure detection in document images. It is evident that the approach with deformable convolutions has outranked the instance segmentation-based approach. This is because of the dynamic receptive field that takes care of the figures having various scales and aspect ratios in the document images. The results also entail that instead of providing raw images to the deep neural network, transforming images through traditional document image analysis methods can yield better results.

5.7. Evaluations for Formula Detection

The performance assessment between the two novel approaches is explained in Table 5. Analogous to the figure detection, the approach with the blend of image transformations and deformable convolutions has out-smarted the other method for formula detection.

While evaluating page object detection systems, it is essential to mention that still there is a room for improvement to come up with deep neural networks that can localize and classify all the page objects present in a document image. So far, we have seen that particular methods or modules are utilized to detect various page objects.

6. Discussion and Conclusion

This paper provides a comprehensive overview of approaches that perform end-to-end graphical element detection in document images. There are different type of challenges:

- First and foremost is that the current state-of-the-art performs better when the network is trained for a single type of graphical object, i.e., only for table or only for formula.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 28 April 2021 doi:10.20944/preprints202104.0739.v1

Table 3: Table detection performance evaluation across various datasets. The double horizontal line divides the employed datasets.

Literature	Year	Dataset	IOU	Precision	Recall	F-Measure	Method
Saha et al. [22]	2019	ICDAR-17 POD	0.6	-	-	0.971	Mask R-CNN
Siddiqui et al. [19]	2018	ICDAR-17 POD	0.6	0.965	0.971	0.968	Deformable Faster R-CNN
Agarwal et al. [30]	2020	ICDAR-13	0.5	1	1	1	CDEC-Net
Saha et al. [22]	2019	ICDAR-13	0.5	0.982	1	0.991	Mask R-CNN
Kavasidis et al. [16]	2018	ICDAR-13	0.5	0.981	0.981	0.981	Fully Convolutional Network
Siddiqui et al. [19]	2018	ICDAR-13	0.5	0.996	0.996	0.996	Deformable FPN
Schreiber et al. [14]	2017	ICDAR-13	0.5	0.974	0.962	0.968	Faster R-CNN
Agarwal et al. [30]	2020	UNLV	0.5	0.960	0.770	0.865	CDeC-Net
Saha et al. [22]	2019	UNLV	0.5	0.946	0.910	0.928	Mask R-CNN
Siddiqui et al. [19]	2018	UNLV	0.5	0.786	0.749	0.767	Deformable FPN
Gilani et al. [15]	2017	UNLV	0.5	0.823	0.907	0.863	Faster R-CNN
Agarwal et al. [30]	2020	ICDAR-2019	0.5	0.987	0.946	0.966	CDeC-Net

Table 4: Figure detection performance comparison.

Literature	Year	Dataset	IOU	Precision	Recall	F-Measure	Method
Younas et al. [24]	2020	ICDAR-17 POD	0.6	0.931	0.913	0.922	Fi-Fo with Deformable Convoltuions
Saha et al. [22]	2019	ICDAR-17 POD	0.6	-	-	0.918	Mask R-CNN

The performance degrades when we train the network for the detection of all types of graphical objects.

- The second important challenge is low inter-class (between different classes) and high intra-class (within the same class) variation. Some of the charts, for example, closely resemble tables and vice-versa.
- The datasets differ significantly from each other. For example, one dataset concentrate fully on tables, and the other dataset concentrate fully on formulas.
- Similarly, different types of datasets have different artifacts resulting from capturing process, such as scanning of the documents.
- The object detection networks are generally huge, and it is not easy to process the images at their original resolution. Therefore, during the downsampling process, some of the minor details are missed.
- In general, we need post-processing in order to obtain suitable results. To date, all of the state-of-the-art methods perform some post-processing.

Because of the challenges mentioned above, one may conclude that there are different types of datasets, and standardization is needed with diversity to tune the methods towards generic graphical object detection in document images. Moreover, the development of methods tailored only for graphical page object detection in document images can significantly improve the performance. This work is one effort to unify the performance of the deep neural network architectures for most renowned datasets. The work itself does not only provide an overview of the methods but also discusses the associated challenges with promising future directions (next section).

Table 5: Formula detection performance comparison.

Literature	Year	Dataset	IOU	Precision	Recall	F-Measure	Method
Younas et al. [24]	2020	ICDAR-17 POD	0.6	0.957	0.952	0.954	Fi-Fo with Deformable Convoltuions
Saha et al. [22]	2019	ICDAR-17 POD	0.6	-	-	0.924	Mask R-CNN

7. Future work

There are many possibilities to explore in order to improve the performance of graphical page object detection in document images. In general, recently proposed novel neural network architectures for object detection [64,65] can improve performance of graphical page object detection systems. The second promising direction is the multimodal processing of the graphical objects. In the case of graphical page object detection, multimodal processing, in the simplest form, is the processing of image information and text information together. An example of such a case is when a figure is categorized as a table and vice versa; the text information can be beneficial. The table is the most complicated object among all the graphical page objects. To improve the performance further, another promising path to explore is the localization of individual columns and rows of the specified tables. Furthermore, identifying headers of the table can significantly help to understand the table's inner structure.

References

1. Chen, Jin, and Daniel Lopresti. "Table detection in noisy off-line handwritten documents." In 2011 International Conference on Document Analysis and Recognition, pp. 399-403. IEEE, 2011.
2. Tupaj, Scott, Zhongwen Shi, C. Hwa Chang, and Hassan Alam. "Extracting tabular information from text files." EECS Department, Tufts University, Medford, USA (1996).
3. Fang, Jing, Liangcai Gao, Kun Bai, Ruiheng Qiu, Xin Tao, and Zhi Tang. "A table detection method for multipage pdf documents via visual seperators and tabular structures." In 2011 International Conference on Document Analysis and Recognition, pp. 779-783. IEEE, 2011.
4. Shafait, Faisal, and Ray Smith. "Table detection in heterogeneous documents." In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 65-72. 2010.
5. Kieninger, Thomas, and Andreas Dengel. "A paper-to-HTML table converting system." In Proceedings of document analysis systems (DAS), vol. 98, pp. 356-365. 1998.
6. Kieninger, Thomas, and Andreas Dengel. "Table recognition and labeling using intrinsic layout features." In International conference on advances in pattern recognition, pp. 307-316. Springer, London, 1999.
7. Kieninger, Thomas, and Andreas Dengel. "Applying the T-RECS table recognition system to the business letter domain." In Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 518-522. IEEE, 2001.
8. Wangt, Yalin, I. T. Phillipst, and Robert Haralick. "Automatic table ground truth generation and a background-analysis-based table structure extraction method." In Proceedings of Sixth International Conference on Document Analysis and Recognition, pp. 528-532. IEEE, 2001.
9. Hu, Jianying, Ramanujan S. Kashi, Daniel P. Lopresti, and Gordon Wilfong. "Medium-independent table detection." In Document Recognition and Retrieval VII, vol. 3967, pp. 291-302. International Society for Optics and Photonics, 1999.
10. Shafait, Faisal, and Ray Smith. "Table detection in heterogeneous documents." In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 65-72. 2010.
11. Gatos, Basilios, Dimitrios Danatsas, Ioannis Pratikakis, and Stavros J. Perantonis. "Automatic table detection in document images." In International Conference on Pattern Recognition and Image Analysis, pp. 609-618. Springer, Berlin, Heidelberg, 2005.
12. e Silva, Ana Costa. "Learning rich hidden markov models in document analysis: Table location." In 2009 10th International Conference on Document Analysis and Recognition, pp. 843-847. IEEE, 2009.
13. Hao, Leipeng, Liangcai Gao, Xiaohan Yi, and Zhi Tang. "A table detection method for pdf documents based on convolutional neural networks." In 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 287-292. IEEE, 2016.
14. Schreiber, Sebastian, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. "Deepdesrt: Deep learning for detection and structure recognition of tables in document images." In 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol. 1, pp. 1162-1167. IEEE, 2017.
15. Gilani, Azka, Shah Rukh Qasim, Imran Malik, and Faisal Shafait. "Table detection using deep learning." In 2017 14th IAPR international conference on document analysis and recognition (ICDAR), vol. 1, pp. 771-776. IEEE, 2017.

16. Kavasidis, Isaak, Carmelo Pino, Simone Palazzo, Francesco Rundo, Daniela Giordano, P. Messina, and Concetto Spampinato. "A saliency-based convolutional neural network for table and chart detection in digitized documents." In International Conference on Image Analysis and Processing, pp. 292-302. Springer, Cham, 2019.
17. Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In European conference on computer vision, pp. 818-833. Springer, Cham, 2014.
18. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
19. Siddiqui, Shoaib Ahmed, Muhammad Imran Malik, Stefan Agne, Andreas Dengel, and Sheraz Ahmed. "Decnt: Deep deformable cnn for table detection." IEEE Access 6 (2018): 74151-74161.
20. Vo, Nguyen D., Khanh Nguyen, Tam V. Nguyen, and Khang Nguyen. "Ensemble of deep object detectors for page object detection." In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication, pp. 1-6. 2018.
21. Gao, Liangcai, Xiaohan Yi, Zhuoren Jiang, Leipeng Hao, and Zhi Tang. "ICDAR2017 competition on page object detection." In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 1417-1422. IEEE, 2017.
22. Saha, Ranajit, Ajoy Mondal, and C. V. Jawahar. "Graphical object detection in document images." In 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 51-58. IEEE, 2019.
23. Younas, Junaid, Syed Tahseen Raza Rizvi, Muhammad Imran Malik, Faisal Shafait, Paul Lukowicz, and Sheraz Ahmed. "FFD: Figure and formula detection from document images." In 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1-7. IEEE, 2019.
24. Younas, Junaid, Shoaib Ahmed Siddiqui, Mohsin Munir, Muhammad Imran Malik, Faisal Shafait, Paul Lukowicz, and Sheraz Ahmed. "Fi-Fo Detector: Figure and Formula Detection Using Deformable Networks." Applied Sciences 10, no. 18 (2020): 6460.
25. Chen, Jin, and Daniel Lopresti. "Table detection in noisy off-line handwritten documents." In 2011 International Conference on Document Analysis and Recognition, pp. 399-403. IEEE, 2011.
26. Tupaj, Scott, Zhongwen Shi, C. Hwa Chang, and Hassan Alam. "Extracting tabular information from text files." EECS Department, Tufts University, Medford, USA (1996).
27. Fang, Jing, Liangcai Gao, Kun Bai, Ruiheng Qiu, Xin Tao, and Zhi Tang. "A table detection method for multipage pdf documents via visual seperators and tabular structures." In 2011 International Conference on Document Analysis and Recognition, pp. 779-783. IEEE, 2011.
28. Shafait, Faisal, and Ray Smith. "Table detection in heterogeneous documents." In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 65-72. 2010.
29. Hao, Leipeng, Liangcai Gao, Xiaohan Yi, and Zhi Tang. "A table detection method for pdf documents based on convolutional neural networks." In 2016 12th IAPR Workshop on Document Analysis Systems (DAS), pp. 287-292. IEEE, 2016.
30. Agarwal, Madhav, Ajoy Mondal, and C. V. Jawahar. "CDeC-Net: Composite Deformable Cascade Network for Table Detection in Document Images." arXiv preprint arXiv:2008.10831 (2020).
31. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440. 2015.
32. Yi, Xiaohan, Liangcai Gao, Yuan Liao, Xiaode Zhang, Runtao Liu, and Zhuoren Jiang. "CNN based page object detection in document images." In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 230-235. IEEE, 2017.
33. Gao, Liangcai, Yilun Huang, Hervé Déjean, Jean-Luc Meunier, Qinqin Yan, Yu Fang, Florian Kleber, and Eva Lang. "ICDAR 2019 competition on table detection and recognition (cTDaR)." In 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1510-1515. IEEE, 2019.
34. Fang, Jing, Xin Tao, Zhi Tang, Ruiheng Qiu, and Ying Liu. "Dataset, ground-truth and performance metrics for table detection evaluation." In 2012 10th IAPR International Workshop on Document Analysis Systems, pp. 445-449. IEEE, 2012.
35. Li, Minghao, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. "Tablebank: Table benchmark for image-based table detection and recognition." In Proceedings of The 12th Language Resources and Evaluation Conference, pp. 1918-1925. 2020.
36. Dai, Jifeng, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. "Deformable convolutional networks." In Proceedings of the IEEE international conference on computer vision, pp. 764-773. 2017.
37. Siegel, Noah, Nicholas Lourie, Russell Power, and Waleed Ammar. "Extracting scientific figures with distantly supervised neural networks." In Proceedings of the 18th ACM/IEEE on joint conference on digital libraries, pp. 223-232. 2018.
38. Mondal, Ajoy, Peter Lipps, and C. V. Jawahar. "IIIT-AR-13K: a new dataset for graphical object detection in documents." In International Workshop on Document Analysis Systems, pp. 216-230. Springer, Cham, 2020.
39. Fukushima, Kunihiko. "Neocognitron." Scholarpedia 2, no. 1 (2007): 1717.
40. Li, Minghao, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. "Docbank: A benchmark dataset for document layout analysis." arXiv preprint arXiv:2006.01038 (2020).
41. Göbel, Max, Tamir Hassan, Ermelinda Oro, and Giorgio Orsi. "ICDAR 2013 table competition." In 2013 12th International Conference on Document Analysis and Recognition, pp. 1449-1453. IEEE, 2013.

42. Shahab, Asif, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. "An open approach towards the benchmarking of table structure recognition systems." In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 113-120. 2010.
43. Zhong, Xu, Jianbin Tang, and Antonio Jimeno Yepes. "Publaynet: largest dataset ever for document layout analysis." In 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1015-1022. IEEE, 2019.
44. Mori, Shunji, Ching Y. Suen, and Kazuhiko Yamamoto. "Historical review of OCR research and development." Proceedings of the IEEE 80, no. 7 (1992): 1029-1058.
45. Breuel, Thomas M. "The OCRopus open source OCR system." Document recognition and retrieval XV. Vol. 6815. International Society for Optics and Photonics, 2008.
46. Hashmi, Khurram Azeem, Rakshith Bymana Ponnappa, Syed Saqib Bukhari, Martin Jenckel, and Andreas Dengel. "Feedback Learning: Automating the Process of Correcting and Completing the Extracted Information." In 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW), vol. 5, pp. 116-121. IEEE, 2019.
47. Pondenkandath, Vinaychandran, Mathias Seuret, Rolf Ingold, Muhammad Zeshan Afzal, and Marcus Liwicki. "Exploiting state-of-the-art deep learning methods for document image analysis." In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 5, pp. 30-35. IEEE, 2017.
48. Rice, Stephen V., Frank R. Jenkins, and Thomas A. Nartker. The fourth annual test of OCR accuracy. Vol. 3. Technical Report 95, 1995.
49. He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn. in Proceedings of the IEEE international conference on computer vision." (2017): 2961-2969.
50. Ren, Shaoqing, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks." arXiv preprint arXiv:1506.01497 (2015).
51. Cai, Zhaowei, and Nuno Vasconcelos. "Cascade R-CNN: high quality object detection and instance segmentation." IEEE transactions on pattern analysis and machine intelligence (2019).
52. Smith, R. (2007, September). An overview of the Tesseract OCR engine. In Ninth international conference on document analysis and recognition (ICDAR 2007) (Vol. 2, pp. 629-633). IEEE.
53. Girshick, Ross. "Fast r-cnn." In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448. 2015.
54. Everingham, Mark, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes (voc) challenge." International journal of computer vision 88, no. 2 (2010): 303-338.
55. Breu, Heinz, Joseph Gil, David Kirkpatrick, and Michael Werman. "Linear time Euclidean distance transform algorithms." IEEE Transactions on Pattern Analysis and Machine Intelligence 17, no. 5 (1995): 529-533.
56. Fabbri, Ricardo, Luciano Da F. Costa, Julio C. Torelli, and Odemir M. Bruno. "2D Euclidean distance transform algorithms: A comparative survey." ACM Computing Surveys (CSUR) 40, no. 1 (2008): 1-44.
57. Ragnemalm, Ingemar. "The Euclidean distance transform in arbitrary dimensions." Pattern Recognition Letters 14, no. 11 (1993): 883-888.
58. Shahab, Asif, Faisal Shafait, Thomas Kieninger, and Andreas Dengel. "An open approach towards the benchmarking of table structure recognition systems." In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 113-120. 2010.
59. Neubeck, Alexander, and Luc Van Gool. "Efficient non-maximum suppression." In 18th International Conference on Pattern Recognition (ICPR'06), vol. 3, pp. 850-855. IEEE, 2006.
60. Krähenbühl, Philipp, and Vladlen Koltun. "Efficient inference in fully connected crfs with gaussian edge potentials." Advances in neural information processing systems 24 (2011): 109-117.
61. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Identity mappings in deep residual networks." In European conference on computer vision, pp. 630-645. Springer, Cham, 2016.
62. Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft coco: Common objects in context." In European conference on computer vision, pp. 740-755. Springer, Cham, 2014.
63. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440. 2015.
64. Chen, Kai, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng et al. "Hybrid task cascade for instance segmentation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4974-4983. 2019.
65. Liu, Ze, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin transformer: Hierarchical vision transformer using shifted windows." arXiv preprint arXiv:2103.14030 (2021).