Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 44 No. 5 (2023)

Image Segmentation And Data Abstraction Using Deep Learning

[1] Niraj Muttur, [2] Rishikesh Patidar, [3] Prashanth Kumar S, [4] Smitha N
[1]-[4] CMR Institute of Technology, Department of Computer Science & Engineering, Bengaluru, India

Abstract—Text recognition in pictures is a very important task in a lot of real-world applications. The growing usage of invoices has created an unnecessary financial industry's labour and physical resources. Localising text from pictures taken with a camera is now a growing demand for modern IT-enabled services. Most of today's text localization techniques are sensitive to the features of text such as colour, size, and style, and also to the cluttered background. Deep Learning approaches help in extracting the characteristics of each frame.

Index Terms —Character segmentation, deep learning, CNN, text recognition, text localization, OCR.

I. Introduction

The amount of unstructured data is now growing every day. Almost 95% of business organisations struggle to effectively analyse unstructured documents. Millions of dollars will be spent on the analysis and integration of this data into the enterprise's information systems. [1]. Government and business departments often require the capture of data from printed documents, however entering data using a computer or any other device is subject to human errors and is a tedious task. Over the past several years, a variety of technologies have been available for optical word detection in photographs taken using various equipment. Bills and other official documents, as well as identification documents, often go through a complicated process. The identification process can be split into several steps, such as identifying and locating documents, retrieving and rectifying areas, field segmentation, field recognition, language template post-processing, and acquisition of results [2]. In the past, paper banknotes have been used to store data about financial transactions. Paper invoices are traditionally used for the storage of financial transaction information. So there's a desire for extracting and storing these statistics from the payments automatically. It is therefore necessary to automatically extract and store this information from invoices. It is possible to do so using OCR.Tools like OpenCV and Tesseract OCR Engine were used to develop the above-noted system[3].The concept of processing and extracting textual content from a photograph is attractive enough.It targets to extract facts from the invoices within the photograph, right here that specialises in payments and invoices only, which could later assist us to locate what humans select nowadays within the marketplace and the way their picks and likings range over geography, time, etc[4].

The suggested system may batch-scan invoices using scanners, automatically extract identified information from invoices, and save identified information at the appropriate spots in the Excel table after saving the invoices. The system is made up of an input module, a pre-processing module, a positioning module, a segment module, an identification module, a judgement and change module, and more, depending on its role. In this document, we first perform a search and analysis of each module and, based on the search results, we determine how invoices are accounted for as follows: 1) In this method, first off normalise the bill; 2) Here, using the projection technique to separate characters and the template fit to find data; 3) CNN is brought in for character identification; 4) Finally, these above-mentioned results are used to alter the modules, which can determine if there is an error based on the relationship between the invoice information, and perform a manual change when it occurs [5].

II. Literature Review

An approach primarily based on matching templates is proposed for collecting invoice data. The whole approach consists of inputting the authentic photo, photo preprocessing, template matching, optical character recognition, and data exporting.

- TEXT RECOGNITION AND EXTRACTION

A CNN template can be created and formed based on the printed text contained in the invoices and this template can be used to find text from other invoices with a similar font. The Tesseract OCR engine was used for retrieving text from processed invoice images. For OCR, text localization, character segmentation, and character recognition must be compared. All that is completed with the aid of using the Tesseract OCR engine. It can extract strains of textual content with phrase and line segmentation with an equal layout i.e. the phrases seem equal in order as within the images. Tesseract OCR engine presents very excessive accuracy while running with revealed textual content in place of handwritten textual content[3].

- EARLY SCANNERS

The first driving force behind the handwritten classification of texts has been the numerical classification of postal mail. Jacob Rabinow's first mail drives included scanning hardware and wired logic to recognize single-spacing fonts. Allum et. al progressed this by making an advanced scanner that allowed for greater versions of how the textual content was written in addition to encoding the facts onto a barcode that became published directly on the letter [6].

- TO THE DIGITAL AGE

The first important part of the OCR software was invented by Ray Kurzweil in 1974 as the software made it possible to recognize any font. This software program used an extra advanced use of the matrix method (sample matching). Basically, this will examine the character bitmaps of the model with the bitmaps of the person being studied and could examine them to decide what person it is. The drawback becomes that software becomes sensitive to versions in dimensioning and differences between the ways each individual writes. To improve at templating, the OCR software program began using characteristic extraction instead of templating. For each individual, the software can search for features such as projection histograms, areas, and geometrical moments [6].

III. COMPARISON OF VARIOUS METHODS

In accordance with Table I , various documents are reviewed where Parul Sahara and Sanjay B.Dhok have used a variety of print and manuscript databases with the greatest segmentation and recognition rates, respectively, of 98.86% and 99.84%, are attained [1]. Yulia s. have used convolutional neural networks (CNN) with different architectures to form models that can with precision words [6] Hassan El Bahi, and Abdelkarim Zatni used the OCR data set on ICDAR2015 smartphone documents [7]. Guillaume et al ., created the private data set with around 10,000 lines. This is a mobile, generic, semi-synthetic data set that can be used to make decisions very quickly [8].

Table I. Comparison Of Various Methods

Model	Paper	Datasets	Methods	Results
Artificial Neural Network and Dynamic Programming	YULIA S. CHERNYSHOVA 1,2, ALEXANDER V. SHESHKUS 1,2, VLADIMIR V. ARLAZAROV 3,4	MIDV-500 Census 1961	The solution proposed is based on artificial neural networks (ANN) and dynamic programming rather than using image processing methods.s for the segmentation step or end-to-end ANN.	The proposed method significantly exceeds the algorithmic method implemented in Tesseract 3.05, the LSTM method (Tesseract 4.00).

Convolutional Neural Network (CNN)	Batuhan Balci, Dan Saadati, Dan Shiferaw	IAM Handwriting Dataset	The two principal approaches to this task: direct classification of words and segmentation of characters.	Character level classification was most successful due to model architectures for classifying and even taking into account the potential of mistakes.
K-Nearest Neighbour (k-NN) classifier is used, as it has intrinsically zero training time.	Parul Sahare, Sanjay B. Dhok	Different databases containing printed as well as handwritten texts.	Character segmentation algorithm, primary segmentation trajectories are obtained by means of the structural property of the characters, while the superimposed and joined characters are separated with the help of the graphical distance theory. Finally, segmentation results are validated with a high precision Support Vector Machine (SVM) classifier.	The results of the comparative analysis show that the algorithms offered are more efficient compared to other contemporary approaches, where the highest segmentation and recognition rates are 98.86 percent and 99.84 percent, respectively.
Optical Character Recognition (OCR)	Harshit Sidhwa, Sudhanshu Kulshrestha, Sahil Malhotra, Shivani Virmani	Printed or handwritten characters	OpenCV was used .The middle image is then processed using the Tesseract OCR engine, which is a character recognition motor.	Our methodology proves to be very precise tested on various invoice input images
CNN classification and optical character recognition	Yu Weng & Chunlei Xia.	Shui character dataset	Convolution The neural network (CNN) allows learning guided by data.	The proposed CNN method was validated against the results of OCR.

BiLSTM-CRF model	DIPALI BAVISKAR, SWATI AHIRRAO, KETAN KOTECHA	MNIST,MIDV 500, DocRED, ICDAR-2019.Legal contract documents,2003 French news articles	The proposed dataset for multiple-disposition unstructured invoice documents is assessed using a variety of feature extraction techniques.	The findings of this work suggest that it is suitable for the template.
Optical Character Recognition(OCR)	Vedant Kumar, Pratyush Kaware, Pradhuman Singh, Dr. Reena Sonkusare Siddhant Kumar	Bills are collected and added to the database	The image is first processed using OpenCV to remove shadows or watermarks found in the image. With longer bill times, using an image bifurcation process.	The system shows us that technologies such as optical character recognition can be used to invoice information correctly.

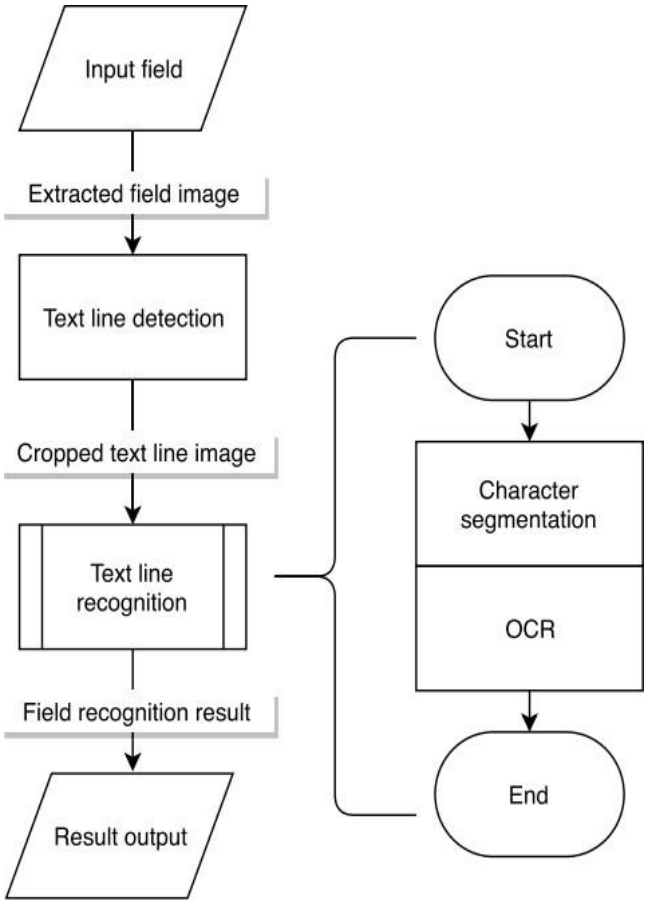

Figure 1: Flow Chart

IV. PROPOSED ARCHITECTURE

Several computer vision techniques were used to identify the table structure of a PDF or image file and using two kernels an attempt was made to detect the edge. 1. Kernel for detecting horizontal lines 2. Kernel for detecting vertical lines. Since structure extraction is currently one of the most important areas of research in deep learning, we decided to use deep neural networks as a use case. This use case is divided into three different sections. 1. Region detection using deep learning techniques, 2. Text extraction from the detected area using the OCR tool. 3. Implement text analysis to identify the relationships between the extracted texts and place them in the repository as shown in Figures 1 .

- Image pre-processing:

Images of the training and testing process are processed here. First, convert the PDF invoice to JPG at (600x600x3) and 300 DPI, then use the various pre-processing techniques. Once all images have been collected and processed, they will be analysed for training using a deep learning model (Yolov5 is used).

- Structure Extraction Using Deep Learning:

The extraction is divided into two categories described below.

- Detection Modelling:

Pass the dataset to the recognition model after it is prepared so that it can determine the table, paragraph, or form of the input image. We are currently using Yolov5.

- Text Extraction:

Text extraction is the pipeline's following step. Tesseract-OCR, an open-source OCR program, is now being used to extract text from the detected region. One of the most crucial challenges in the field of object recognition is image labelling. We manually labelled 1000 photographs in three categories—paragraph, table, and form—using the GUI-based labelling program labelling. The Yolo format is used for all annotations, and they are all stored as TEXT files in a different directory as shown in Figure 2.

Figure 2: Invoice Data Segmentation.

In Figure 2 the model will give the segmented image for the input image and the bounding boxes indicate the identified paragraph and table classes in the input image [10].

OCR recognizes the text content of an image and converts the image into machine-coded text that can be processed by a computer[11-13]. The following is the OCR procedure. The image is scanned to produce a bitmap, which is a matrix of black and white dots. For better accuracy, the image should be preprocessed with brightness and contrast adjusted. Then use a segmentation algorithm to find the area of interest in images where the text is located. Now we can further divide the area of interest into lines, words, and letters. Later comparison and recognition algorithms could be used to match the characters [14][15].

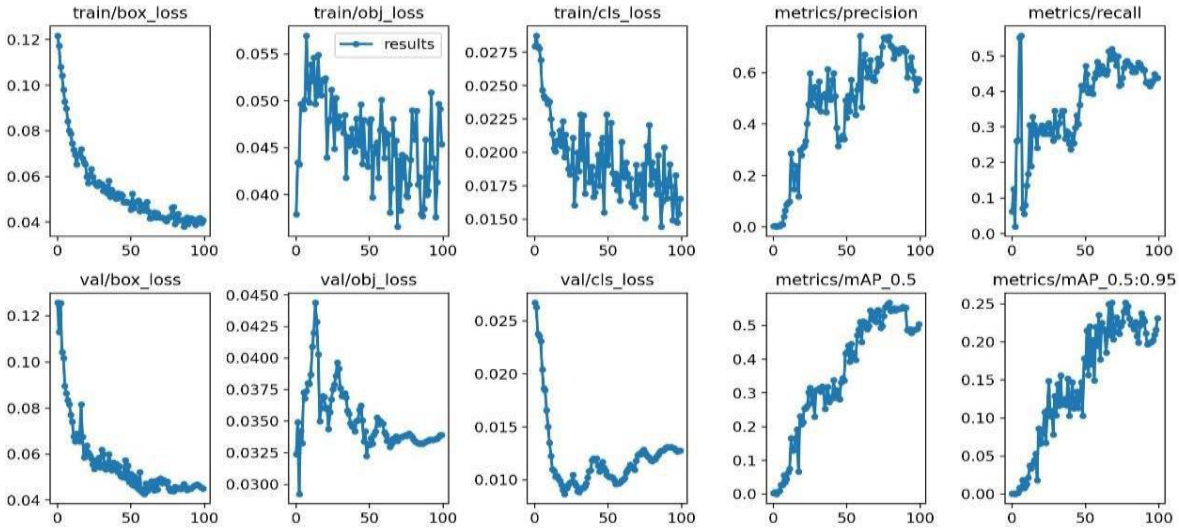

Figure 3: Resulting matrix graph

In order to continue, find the Tesseract execution file, copy the output and specify the location of the .exe file. , upload the required images to notebook storage. In this case, the image is called Invoice_46_brightness2.jpg [16].

The results.csv file will be created which contains the summary of accuracy and losses achieved at each epoch as shown in Figure 3.

V. CONCLUSION

In this document, a detailed review is done of the text recognition and the system of extracting text from images (invoice images), where most of the work used OCR to extract text from invoice pictures with a simple background and similar template. But for loud background bills, watermarks, and also with several layouts, it will be difficult. To enhance performance, we can expand the size of the dataset with multi-format invoice images and automate invoice annotation.

We investigated numerous state-of-the-art deep neural network models, chose a few of them as application cases, and carried out all necessary data preprocessing for training operations. To help train models, we have produced a real dataset. In order to create graphical models that more effectively handle and enhance the relationships between table headers, rows, and cells, we are looking at the use of graph convolutional neural networks in the future.

As you can see, automated invoice processing is not just for bills. This concept can be applied to the financial and banking sectors, as well as any other industry with a lot of paperwork.

Cost reduction: While extracting data from calculations using traditional methods, we need to develop rule-based engines and constantly change them as the data becomes more volatile. This increases the implementation and other operational costs of processing the invoice. Deep learning data extraction processes help increase system efficiency, reduce errors, and generate significant profits in a short time frame.

Efficient Process Management: We use traditional methods to extract data from calculations, but we need to develop rule-based engines and constantly change them as the data becomes more volatile. This

increases the implementation and other operational costs of invoice processing. Deep learning data extraction processes help increase system efficiency, reduce errors, and generate significant profits in a short time frame.

VI. Future Work

Developing an AI-powered tool stands as another substantial research initiative, poised to streamline the extraction of vital data fields and minimise the need for manual efforts in the acquisition and auditing of unstructured documents, all while ensuring originality.Unstructured handwritten papers, such as cashier receipts and doctor's notes, contain valuable information for extracting and analysing crucial areas. The development of a model for on-the-spot retrieval of printed data and coupled unstructured handwriting information will enhance this study.

In summary, Many extraction techniques have been developed to search for relevant information. Therefore, successful implementation of text extraction from invoice images in any organisation requires identifying business goals and analysing data accessible from both open source and private datasets.

References

[1] Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches.DIPALI BAVISKAR, SWATI AHIRRAO and KETAN KOTECHA.Digital Object Identifier 10.1109/ACCESS.2021.3096739

[2] Two-Step CNN Framework for Text Line Recognition in Camera-Captured Images.YULIA S. CHERNYSHOVA,ALEXANDER V. SHESHKUS and VLADIMIR V. ARLAZAROV.Digital Object Identifier 10.1109/ACCESS.2020.2974051.This work was supported in part by the RFBR According to the Research Project 17-29-03170 and Project 17-29-03236.

[3] Extraction of information from bill receipts using optical character recognition.Vedant Kumar, Pratyush Kaware, Pradhuman Singh, Dr.Reena Sonkusare, and Siddhant Kumar.Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020). IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1-7281-5461-9.

[4] Text Extraction from Bills and InvoicesHarshit Sidhwa, Sudhanshu Kulshrestha, Sahil Malhotra, Shivani Virmani.International Conference on Advances in Computing, Communication Control and Networking (ICACCCN2018)

[5] An Invoice Recognition System Using Deep Learning.Shaoqing Shi,Chao Cui and Yong Xiao.2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS)

[6] Handwritten Text Recognition using Deep Learning.Batuhan Balci,Dan Saadati and Dan Shiferaw.

[7] Text recognition in document images obtained by a smartphone based on deep convolutional and recurrent neural network Hassan El Bahi and Abdelkarim Zatni

[8] Handwritten text line segmentation using Fully Convolutional Network.2017 14th IAPR International Conference on Document Analysis and Recognition.

[9] HIDDEN MARKOV MODEL-BASED OPTICAL CHARACTER RECOGNITION IN THE PRESENCE OF DETERMINISTIC TRANSFORMATIONS.OSCAR E. AGAZZI and SHYH-SHIAW KuoSignal Processing Research Department, AT&T Bell Laboratories, and Murray Hill, NJ 07974, U.S.A.

[10] Text segmentation using superpixel clustering ISSN.Yuanping Zhu1, Kuang Zhang.

[11] An Independent Character Recognizer for Distantly Acquired Mobile Phone Text Images

[12] An Independent Character Recognizer for Distantly Acquired Mobile Phone Text Images.Binh Quang Long Mai, Tue Huu Huynh, Anh Dong Doan.

[13] OPTICAL CHARACTER RECOGNITION WITHOUT SEGMENTATION.Mehmet Ali Ozdil - Fatos T, Yarman Vural

[14] Scene text detection and recognition: recent advances and future trends.Yingying ZHU, Cong YAO, Xiang BAI

[15] Text Detection, Tracking and Recognition in Video: A Comprehensive Survey.Xu-Cheng Yin, Senior Member, IEEE, Ze-Yu Zuo, Shu Tian, and Cheng-Lin Liu

[16]https://medium.com/analytics-vidhya/invoice-information-extraction-using-ocr-and-deep-learning-b79464f54d69