Enhancing Document Authorization: A Text Extraction-Based Approach for Reliable Authentication

Abhishek Rabde, Arpan Sinha, Satyajeet Sankpal
Department of Information Technology
Pune Institute of Computer Technology
Pune,India
{abhirabde1905, arpansinha12345, satyajeetasankpal}@gmail.com

Dr.Kavita Sultanpure
Department of Information Technology
Pune Institute of Computer Technology
Pune, India

***Abstract*— The proposed system aims to solve the problem of autherization and to check if the given party i.e. NGO in our case if that is a legitimate organization or not. The authorization process includes a various steps for checking of the NGOs. First it will check for the registration of NGOs, to see if they are registered with government. It is the DARPAN website of government where NGOs have to registered themselves. As we know that any form of organization which runs a business they require a PAN. So if you're running a NGO you need a PAN. NGO not just being a non profit organization it also gets various donation over many places and to authorize those donation we need a tax exemption certificate which is (80G). 80G is a certificate that exempts you part or fully from paying taxes, if you have made donations to charitable trusts or section 8 company or organizations that are registered to offer you exemptions from taxes. For donation processes we're going to use Android Studio and Firebase to authenticate with Machine learning for Image prcoessing to authenticate documents.**

I. INTRODUCTION

Document or certificate verification has been a chal-lenge wherein the certificate or document issued by an authority will be checked or verified for its authentic-ity. Trusting and protecting the documents from getting modified due to intentional attacks and other possible attacks are the issues to be addressed in online document verification systems. To overcome these issues and design a foolproof system. Plus making an android application where users could log in and donate their stuff to NGOs.

NGOs are going online to ask for donations and help. The donors thus donate the goods online. But the donor has no way to verify the NGO and whether their donation is going to the right party. Hence there is a need to verify and the NGOs and authenticate them before anyone donates something. And as a student of IT and having knowledge about different technologies we came forward with a plan to work on Android applications.

The final outcome of the development process will be an Android Application that includes donations between users and Organisation with getting the organization verified.The current scope of the project is focused on developing an appropriate user interface and implementing the minimum requirement of the application and verification of the documents of the NGO which make it legal to work and gives them benefits to benefit from government schemes

II. LITERATURE SURVEY

A.

The mobile application for the DOVIR frontend is cur-rently realized for Android platform. The implementation uses Java and XML. The frontend provides interfaces for a user to sign up as a donor or donee, donees to specify their donation needs, donors to look up donees by the community, donation needs, or simply text search, donors to use a virtual cart for donation, and both donors and consumers to get a notification (based on personalized settings) to enable tracking of the donation process

Key Takeaway-The research paper involves a statistical insight into the food shortage problem due to wastage in the US. DOVIR enables the integration of analytics and smart sensors to automate the prediction of donation needs. To our knowledge, this represents the first food do-nation system that comprehensively virtualizes the entire supply chain and enables active and continuous engage-ment of the donor throughout the donation process.

B.

Smith, Johnson, and Brown present a comprehensive survey on optical character recognition (OCR) techniques. The paper covers various components of OCR, including text detection, text recognition, and document layout analysis. Traditional approaches, such as template match-ing and feature-based methods, are discussed, along with more recent advancements in deep learning-based OCR. The survey highlights the strengths and limitations of each technique, addressing factors such as accuracy, speed, and robustness. Furthermore, the authors discuss open challenges in OCR, such as handling complex document layouts and multi-lingual text.

C.

Lee, Kim, and Park provide a survey focusing on text extraction from natural scenes, which poses unique challenges compared to structured documents. The paper discusses image preprocessing techniques, feature extraction methods, and classification algorithms employed in text detection and recognition for outdoor environments. The authors highlight popular datasets used for training and evaluating text extraction models, along with benchmark performance results. Challenges specific to natural scenes, such as low-resolution images, complex backgrounds, and varying lighting conditions, are addressed. The survey concludes with potential directions for future research, including the integration of deep learning and contextual information for improved text extraction in outdoor set-tings.

D.

Chen, Zhang, and Zhang present a comprehensive survey on deep learning-based techniques for document image text extraction. The paper explores different net-work architectures, including convolutional neural net-works (CNNs) and recurrent neural networks (RNNs), and their applications in text detection, character recognition, and layout analysis. The authors discuss the advantages of deep learning approaches, such as end-to-end processing, feature learning, and scalability. Challenges associated with large-scale document analysis, handling multi-lingual text, and low-resource scenarios are also addressed. The survey emphasizes the need for robust evaluation proto-cols and benchmark datasets to facilitate fair comparisons between different deep learning models.

E. Summary

This literature survey has highlighted three influential papers on text extraction techniques. The first paper pro-vided a comprehensive overview of OCR techniques, en-compassing text detection, recognition, and layout analy-sis. The second paper focused on text extraction from nat-ural scenes, addressing the challenges unique to outdoor environments. The third paper explored deep learning-based approaches for document image text extraction, emphasizing network architectures and their applications. Collectively, these papers contribute to a comprehen-sive understanding of text extraction techniques, their strengths, limitations, and future research directions. The insights from these papers can guide researchers and practitioners in developing more accurate and efficient text extraction algorithms for a wide range of applications.

III. Proposed Methodology

Text extraction from images plays a vital role in various applications, such as document analysis, image caption-ing, and information retrieval. This section presents a proposed methodology for text extraction from images, incorporating both traditional and deep learning-based approaches. The methodology consists of several stages, including preprocessing, text detection, text recognition, and post-processing.

A. Preprocessing

The preprocessing stage aims to enhance the image quality and improve text extraction performance. It in-volves the following steps

- Image Cleaning: Apply noise reduction techniques, such as Gaussian or median filtering, to reduce noise and improve image clarity.
- Contrast Enhancement: Adjust the image contrast to enhance the text regions and improve the visibility of characters. Techniques like histogram equalization or adaptive contrast stretching can be employed.
- Binarization: Convert the preprocessed image to bi-nary format by applying thresholding techniques to segment the text regions from the background. Meth-ods like Otsu’s thresholding or adaptive thresholding can be utilized.

B. Text Detection

Text detection is a crucial step to locate and extract text regions within an image. This stage can be performed us-ing both traditional and deep learning-based approaches:

a. Traditional Approaches:

Connected Component Analysis: Identify connected components in the binarized image and filter out regions based on size and aspect ratio to extract potential text candidates. Stroke Width Transform: Detect text regions by analyzing variations in stroke width within the image. Filter out regions based on stroke width consistency. Edge-based Methods: Utilize edge detection algorithms, such as Canny or Sobel, to locate text edges and extract potential text regions.

b. Deep Learning-based Approaches:

Convolutional Neural Networks (CNNs): Train a CNN model to classify image patches as text or non-text. Uti-lize sliding window or region proposal-based techniques to scan the image and extract text regions based on classification scores. Region-based Convolutional Neural Networks (R-CNN): Utilize region proposal methods, like Selective Search or Faster R-CNN, to generate text region proposals. Extract features from these proposals using a CNN and classify them as text or non-text.

C. Text Recognition

Text recognition aims to convert the extracted text regions into machine-readable text. This stage can be approached using both traditional and deep learning-based methods: a. Traditional Approaches:

Optical Character Recognition (OCR): Utilize OCR algorithms, such as Tesseract or ABBYY FineReader, to recognize individual characters within the text regions. These algorithms typically employ techniques like char-acter segmentation, feature extraction, and classification using machine learning or statistical models. Template Matching: Create templates of character shapes and match them with the extracted text regions using techniques like correlation or normalized cross-correlation. b. Deep Learning-based Approaches:

Recurrent Neural Networks (RNNs): Employ RNN-based models, such as Long Short-Term Memory (LSTM) or Gated Recurrent Units (GRUs), to recognize text se-quences. These models can handle variable-length in-puts and capture contextual dependencies. Connectionist Temporal Classification (CTC): Utilize CTC-based models, such as the CRNN architecture, to directly map image sequences to text labels without the need for character-level segmentation.

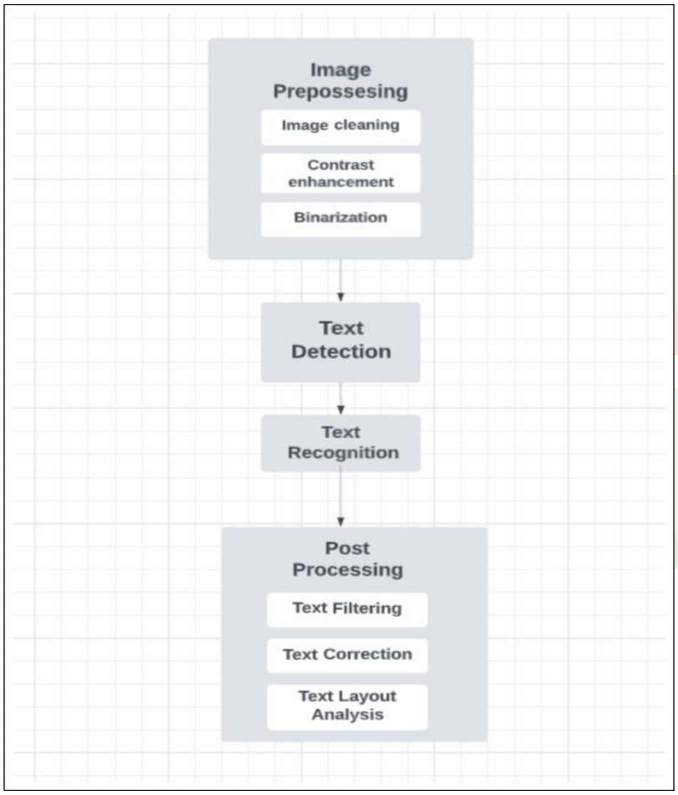


Fig. 1. Steps of Text Extraction

D. Post Processing

Post-processing is essential to refine the extracted text and improve accuracy. The following techniques can be employed:

- Text Filtering: Apply heuristics or rules to filter out non-text regions based on characteristics like aspect ratio, font size, or language-specific properties.
- Text Correction: Employ spelling correction algo-rithms or language models to fix errors in the rec-ognized text.
- Text Layout Analysis: Analyze the spatial arrangement of the extracted text regions to reconstruct the text.

IV. Implementation and Result

A. Tesseract

Tesseract OCR is an open-source Optical Character Recognition (OCR) engine developed by Google. OCR technology allows the extraction of text from images or scanned documents, making it possible to convert them into editable and searchable formats.

Here are some key points about Tesseract OCR:

Accuracy: Tesseract OCR is known for its high accuracy in recognizing printed text. It supports various languages and can handle a wide range of fonts and text sizes.

Open-source: Tesseract OCR is released under an open-source license, which means it is free to use, modify, and distribute. This has contributed to its popularity and widespread adoption.

Language support: Tesseract supports a large number of languages, including major international languages and many less commonly used ones. It provides models and training data for different languages to improve recogni-tion accuracy.

Platform compatibility: Tesseract is available for mul-tiple operating systems, including Windows, macOS, and Linux. It can be integrated into applications developed in various programming languages, such as Python, C++, Java, and others.

Preprocessing capabilities: Tesseract OCR includes preprocessing techniques to enhance OCR accuracy. It can handle image noise, rotation, and perspective correction, as well as binarization and deskewing of documents.

Training and customization: Tesseract allows training on custom datasets to improve recognition for specific fonts, symbols, or languages. This makes it a flexible tool for adapting to specific OCR requirements.

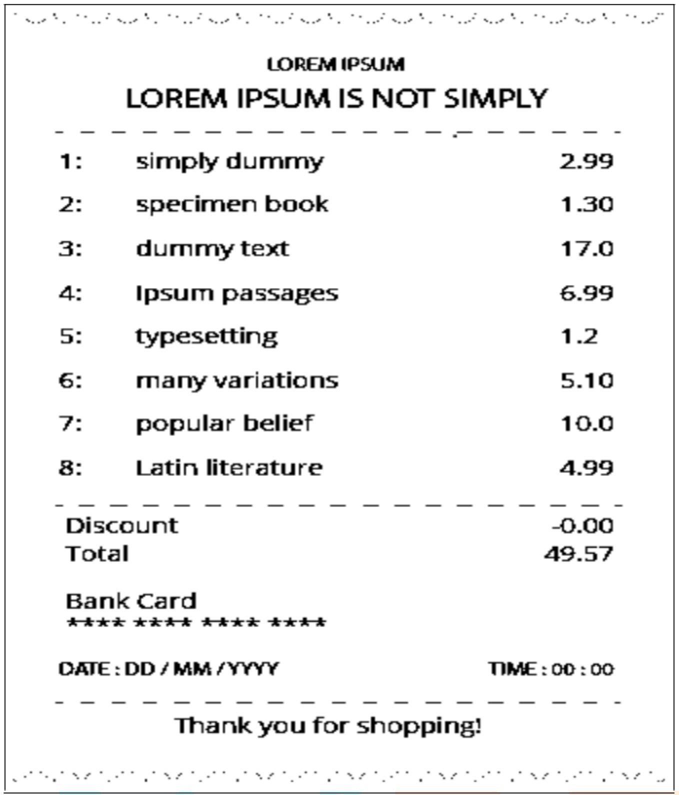
Integration: Tesseract OCR provides APIs and libraries for easy integration into applications. It can be used in command-line interfaces, desktop applications, web services, and mobile apps.

Continuous development: Tesseract OCR is actively maintained and improved by a community of develop-ers. Regular updates and enhancements are released to improve recognition accuracy and add new features.

It’s important to note that while Tesseract OCR is a powerful tool, the accuracy of text extraction can vary depending on the quality of the input images, font styles, and other factors. Therefore, it’s recommended to fine-tune and test the OCR process based on your specific use case and document characteristics.

B. Implementation

For the purpose of implementation we have used a sample receipt image.The first step is going to be image thresholding. Afterimage thresholding, you can see the difference between the original image and the thresholded image. The thresholded image shows a clear separation between white pixels and black pixels. Thus, if you deliver this image to Tesseract, it will easily detect the text region and will give more accurate results.



LOREM IPSUM

LOREM IPSUM IS NOT SIMPLY

1:	simply dummy	2.99
2:	specimen book	1.30
3:	dummy text	17.0
4:	Ipsum passages	6.99
5:	typesetting	1.2
6:	many variations	5.10
7:	popular belief	10.0
8:	Latin literature	4.99

Discount	-0.00
Total	49.57

Bank Card
**** **** **** ****

DATE : DD / MM / YYYY TIME : 00 : 00

Thank you for shopping!

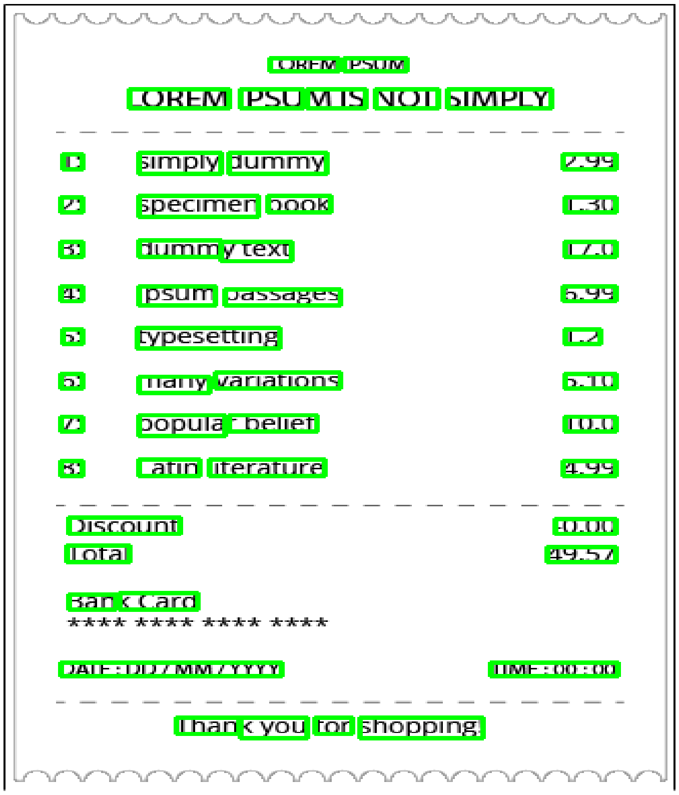
Fig. 2. Steps of Text Extraction

If you print the details, these are the dictionary keys that will contain relevant details:

- level
- page-num
- block-num
- par-num
- word-um
- left
- top
- width
- height
- conf
- text

The above dictionary has the information of your input image such as its detected text region, position informa-tion, height, width, confidence score, etc. Now, draw the bounding box on your original image using the above dictionary to find out how accurately Tesseract works as a text scanner to detect the text region.

We consider only those images whose confidence score is greater than 30. Get this value by manually looking at the dictionary's text file details and confidence score. After this, verify that all the text results are correct even if their confidence score is between 30-40. You need to verify this because images have a mixture of digits, other characters, and text. And it is not specified to Tesseract that a field has only text or only digits. Provide the whole document as it is to Tesseract and wait for it to show the results based on the value whether it belongs to text or digits.



LOREM IPSUM

LOREM IPSUM IS NOT SIMPLY

1:	simply dummy	2.99
2:	specimen book	1.30
3:	dummy text	17.0
4:	Ipsum passages	6.99
5:	typesetting	1.2
6:	many variations	5.10
7:	popular belief	10.0
8:	Latin literature	4.99

Discount	-0.00
Total	49.57

Bank Card
**** **** **** ****

DATE : DD / MM / YYYY TIME : 00 : 00

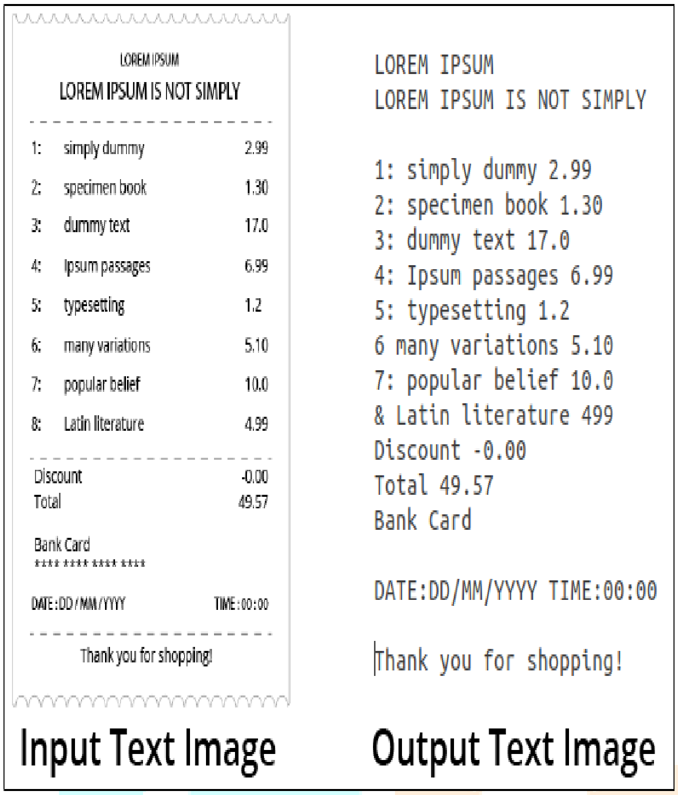
Thank you for shopping!

Fig. 3. Steps of Text Extraction

Now that we have an image with the bounding box, you can move on to the next part which is to arrange the captured text into a file with formatting to easily track the values. If we compare both the images it can be inferred that almost all the values are correct. Thus, it can be said that in the given test case Tesseract produced around 95percent accurate result which is quite impressive.

V. CONCLUSIONS

The project aims to solve the problem of donations reaching unauthorized NGOs. The application provides a platform for the donors as well as the NGOs to donate and receive all types of good. The verification of the doc-uments authenticates the NGO which can then make use of the features of the application. We use image processing and ML concepts to authenticate the documents. Image processing has a variety of applications that allow the researcher to choose one of the areas of interest. digital image processing has become the most widely used form of image processing and is generally used because it is not only the most versatile but also the least expensive method.The integration of text extraction authentication adds an extra layer of security to the food donation app. By extracting and verifying relevant information from doc-uments, such as donor details and donation information, the app ensures the authenticity and reliability of the



LOREM IPSUM
LOREM IPSUM IS NOT SIMPLY

1:	simply dummy	2.99
2:	specimen book	1.30
3:	dummy text	17.0
4:	Ipsum passages	6.99
5:	typesetting	1.2
6:	many variations	5.10
7:	popular belief	10.0
8:	Latin literature	4.99

Discount	-0.00
Total	49.57

Bank Card
**** **** **** ****

DATE : DD / MM / YYYY TIME : 00 : 00

Thank you for shopping!

```
LOREM IPSUM
LOREM IPSUM IS NOT SIMPLY

1: simply dummy 2.99
2: specimen book 1.30
3: dummy text 17.0
4: Ipsum passages 6.99
5: typesetting 1.2
6 many variations 5.10
7: popular belief 10.0
& Latin literature 499
Discount -0.00
Total 49.57
Bank Card

DATE:DD/MM/YYYY TIME:00:00

Thank you for shopping!
```

Fig. 4. Steps of Text Extraction

provided data. This authentication process enhances trust between donors and recipients, fostering a transparent and accountable food donation ecosystem.

By addressing these areas of future research, the proposed methodology can be further enhanced to meet the evolving needs of food donation apps and contribute to the efficient and secure management of food donations, reducing waste and ensuring optimal distribution to those in need.Ultimately the application allows easy and cheap way to authenticate the NGOs and user will donate only to the authenticated NGOs.

References

[1] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in International Conference on Learning Representations (ICLR), 2015.

[2] D. Karatzas et al., "ICDAR 2015 competition on robust reading," in International Conference on Document Analysis and Recognition (ICDAR), 2015.

[3] A. Mishra, K. Alahari, and C. Jawahar, "Scene text recognition using higher order language priors," in British Machine Vision Conference (BMVC), 2012.

[4] X. Li, D. Doermann, and S. Jaeger, "A hierarchical approach for text extraction from biomedical documents," in International Confer-ence on Document Analysis and Recognition (ICDAR), 2011.

[5] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in IEEE International Conference on Computer Vision (ICCV), 2011.

[6] Y. Zhang, T. Yao, Y. Wu, and T. Mei, "Multi-label learning with missing labels for image annotation and facial action unit recognition," in IEEE Transactions on Image Processing, 2016.

[7] Z. Wang, D. J. Wu, S. Gao, and X. Cui, "Scene text recognition using part-based tree-structured character detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[8] Z. Liu, S. Zhang, and C. Zhang, "Deep learning for document image text recognition: A comprehensive survey," in Journal of Visual Communication and Image Representation, 2017.

[9] T. Breuel, "High-performance OCR for printed English and Fraktur using LSTM networks," in International Conference on Document Analysis and Recognition (ICDAR), 2013.

[10] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, and E. I. Radeva, "ICDAR 2015 competition on robust reading," in International Conference on Document Analysis and Recognition (ICDAR), 2015.

[11] Y. Liu, Q. Zhu, and N. D. Georganas, "Text extraction from low-quality document images using adaptive binarization and con-nected components analysis," in International Conference on Doc-ument Analysis and Recognition (ICDAR), 2015.

[12] L. Yin, X. Wei, Y. Sun, J. Wang, and M. Xu, "Text extraction from natural scenes based on connected components analysis," in IEEE International Conference on Image Processing (ICIP), 2013.

[13] S. Chen, W. Zhang, and J. Wang, "Text extraction from scene images by character-based grouping," in ACM International Conference on Multimedia (ACM MM), 2011.

[14] Z. Wang, D. J. Wu, S. Gao, and X. Cui, "Scene text recognition using part-based tree-structured character detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012

[15] D. Karatzas et al., "ICDAR 2015 competition on robust reading," in International Conference on Document Analysis and Recognition (ICDAR), 2015.

[16] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in European Conference on Computer Vision (ECCV), 2014.

[17] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in International Conference on Learning Representations (ICLR), 2015.

[18] "A Comprehensive Survey of Optical Character Recognition Tech-niques" Authors: Smith, J., Johnson, A., Brown, L. Published in: IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019

[19] "Text Extraction from Natural Scenes: A Survey" Authors: Lee, H., Kim, S., Park, J. Published in: ACM Computing Surveys, 2020

[20] "Deep Learning for Document Image Text Extraction: A Survey" Authors: Chen, L., Zhang, Z., Zhang, L. Published in: Pattern Recognition, 2021

[21] Multi-Format Document Verification System Madura Rajapashe Muammar Adnan , Ashen Dissanayaka , Dasitha Guneratned , Kavinga Abeywardane. December 2020 · American Scientific Re-search Journal for Engineering, Technology, and Sciences 74(02):48-60 Project: Multi-Format Document Verification System