Text Extraction & Recognition from Visiting Cards

1st Chandradeep bhatt
CSE Department
GEHU, Dehradun, India
bhattchandradeep@gmail.com

2nd Akash Sayana *CSE Department*
GEHU, Dehradun, India
Akashsayana05@gmail.com

3rd Rahul Chauhan *CSE Department*
GEHU, Dehradun, India
Chauhan14853@gmail.com

4th Teekam Singh
CSE Department GEHU, Dehradun, India
teekamsingh@gmail.com

***Abstract* -** The increasing demand for automation in business card data entry has led to the development of various tools and techniques for optical character recognition (OCR). In this research, I put forward a flask-based Python application that processes input images of visiting or business cards, detects text regions, and recognizes text to extract important information such as name, address, contact number, email, and website. Our proposed system uses a combination of image pre-processing techniques, machine learning algorithms, and text analysis methods to achieve high accuracy in recognizing the text from the input images. The extracted information is then stored in an Excel table for further use. Our system's experimental results are presented in this study on a dataset of business card images and show that it achieves high accuracy in recognizing the text and extracting the relevant information. Our proposed system provides an efficient and effective solution for business card data entry, which can be applied in various industries such as marketing, sales, and customer relationship management.

Keywords – Image pre-processing, Optical Character Recognition, Text Detection, Text Analysis, Named Entity Recognition.

I. INTRODUCTION

Business cards are an essential tool for networking and communication in today's fast-paced business world. They are a convenient and efficient way to exchange contact information and establish professional connections. However, manually entering the information from these cards into a database can be a time-consuming and error-prone task. The traditional method of managing business cards involves manually transcribing contact information into a digital format, which can lead to errors and inefficiencies. As businesses increasingly rely on digital communication, it is becoming more critical to have a reliable and efficient system for managing business card information. This project aims to solve the challenges of managing business cards by creating a web application that automates the process of extracting relevant information from business cards. The web application uses OCR (Optical Character Recognition) to recognize text on the business card images and NLP (Natural Language Processing) techniques such as Named Entity Recognition and Regular Expressions to identify important pieces of information such as name, company, address, contact number, and email address. The importance of accurate information on business cards cannot be overstated. Incorrect or outdated information can lead to missed opportunities, lost business, and damage to reputation. The digitization of business cards offers numerous benefits over the traditional method of manually transcribing information. Digitized business cards can be easily accessed and searched, eliminating the need for physical storage and reducing the risk of loss or damage. Moreover, digitized business cards offer greater flexibility, allowing users to sort and organize information in various ways. In this report, we will discuss the importance of business cards in the business world and the challenges that arise from managing a large number of them. We will also describe the benefits of digitization and explain how OCR and NLP technologies can be used to automate the process of extracting information from business cards. Finally, we will provide a detailed description of the web application we have developed, its features, and its functionality. Overall, this project aims to create a user-friendly and efficient solution for managing business card information. The development of this web application can significantly reduce the time and effort required to manage business cards, leading to improved productivity and a more efficient workflow.

II. RELATED WORK

Previous research in OCR systems for business cards has primarily focused on various techniques such as character fragmentation and region of interest (ROI) detection to extract text from business cards. Shin, H., & Kim, C. proposed a study in 2014[1].The primary objective of this research is to present a hybrid OCR agent that can enhance the effectiveness of business card recognition in mobile environments. The hybrid OCR agent integrates data from different algorithms and recognition engines with learning capabilities to achieve a better recognition rate. Furthermore, the research proposes an Image Processing Method that can be used on mobile cameras to adjust to variations in lighting, exposure axis, and card backgrounds, thereby improving the accuracy of recognition. Shaaban, A. Hussein, E. Shokry, in 2014 describes the creation of an Arabic OCR system for Egyptian ID cards[2]. It extracts, recognizes, and translates data from the image to editable text, utilizing baseline-based segmentation and DCT-based feature extraction techniques. Experimental results demonstrate the system's dependability, with an average processing time of 16 seconds per ID card using Matlab. Xing Bangli in 2015 proposed an electronic business card management system specifically developed for intelligent wristwatches and incorporates various components like touch screen, camera, three-axis gyroscope, gravity sensor, picture character recognition module, Bluetooth module, timekeeping module, and memory chip[3]. The system employs an intelligent key to initiate the management and electronic business card exchange mode. Additionally, shaking the wrist triggers the camera capture mode. The system replaces conventional paper business cards with electronic ones, allowing for efficient and hassle free exchange and management. Moreover, the system utilizes picture character recognition module and Bluetooth technology for accurate and efficient recognition and exchange of business cards.

Madan Kumar, C. and Brindha, M., proposed a study in 2019 in which the author processed business cards' language line by line and used previous stage acknowledgment for each line[4]. The business cards had fifteen fields, including personal details such as first and last name, organization, email, address, office, and mobile numbers. The author then matched the Tesseract-OCR output with suitable logics and stored it in a dictionary with keywords representing each field. Additionally, a list of extracted data separated by newline characters was created.

H. Goodrum , K. Roberts , E. V. Bernstam conducted a study in 2020 which involved analysing the distribution of digitized records in Electronic Health Records (EHRs) and developing a system to categorize them based on their clinical relevance[5]. The researchers used Optical Character Recognition (OCR) to extract text and assessed various text categorization machine learning models. The study aimed to demonstrate the accuracy of text classification systems in classifying scanned documents. The research conducted by Qinggang M., Minglong S., and Haipeng Y. on "A Business Card Recognition System Based on YOLOv3 and CRNN (2021)" introduces an innovative approach to business card recognition that combines two deep learning models: YOLOv3 for object detection and CRNN for text recognition[6]. Their system accurately extracts relevant information from business cards, while also being able to handle various languages and fonts. H.Wonseok, H. Lee, J. Yim, G. Kim, and M. Seo proposed another study in 2021, the authors aimed to simplify the process of information extraction from semi-structured document images by transitioning from a pipeline-based system to an end-to-end model[7]. The study emphasizes the pragmatic obstacles encountered during the implementation of the system in a vast production environment and formulates document IE as a sequence generation problem. The research showcased that a meticulously designed end-to-end IE system can accomplish proficient performance. Recently, M. Cai, J. Wang, Z. Xu, et al. (2022) propose an efficient business card recognition system based on deep learning[8]. Their system uses a deep convolutional neural network architecture to obtain attributes from business card images and accurately recognize text in multiple languages. The proposed system can extract key information such as name, organization, phone number, and email address. The authors introduce a novel data augmentation technique to improve the accuracy of text recognition. Their approach generates synthetic training data from a limited set of real data, which can reduce overfitting and improve generalization performance. Additionally, the authors use a multi-task learning approach that jointly learns text recognition and data correction, which can help correct errors in the recognized text. Nguyen-Trong, Khanh proposed a new study in 2022, study aimed to develop a method for extracting data from Vietnamese ID card pictures, using deep learning-based techniques and image processing[9]. The authors utilized four neural networks, two basic natural language processing techniques, and four datasets, including manual and synthetic datasets, to evaluate their proposed method. Moreover, a microservice-based architecture was put forward by the researchers to deploy the method on a real-time system. The study's remarkable contributions encompass the proposed technique, the utilization of four distinct neural networks and two image processing techniques, and the development of datasets for Vietnamese OCR.

In our project, we focused on image pre-processing techniques to clean the image and remove unwanted lines, such as vertical, horizontal, and diagonal lines, using structuring elements and the Hough Line Transform. Our approach aimed to simplify the image and improve the accuracy of text recognition by removing unnecessary visual noise. By using these techniques, we believe our approach can improve the overall efficiency and accuracy of OCR systems for business cards, which will ultimately benefit users who need to extract information quickly and accurately from business cards. The proposed system is expected to produce accurate and reliable results and will provide a better user experience.

III. CHALLENGES AND PROBLEMS

The process of text detection and recognition from business cards is faced with various challenges and issues. One of the most significant problems is the quality of the input image, which can greatly affect the accuracy of OCR. Factors like low resolution, poor contrast, and other visual disturbances can lead to errors in recognition[10]. Another challenge is the placement of the image, which can result in incomplete or inaccurate text extraction. To overcome these challenges, preprocessing techniques like contrast adjustment, noise reduction, and filtering can be used to improve the image quality before OCR. Machine learning algorithms can also be employed to train the system to recognize various font styles and layouts, while post-processing techniques can correct any errors in recognition. Moreover, techniques like the perspective transform and the biggest contour method can help isolate the region of interest (ROI) and improve the accuracy of text detection and recognition. By addressing these challenges, the system can be designed to accurately extract and recognize text from business cards.

In addition to the techniques mentioned above, machine learning algorithms can be trained on a large dataset of business card images to improve the accuracy and robustness of the OCR system. This approach can help the system recognize different font styles and card designs, even in cases where the image quality is poor or the placement of the image is not ideal. By implementing these techniques and leveraging the power of machine learning, the system can be developed to accurately extract and recognize text from business cards, ultimately improving efficiency and productivity in various industries.

IV. DATASET

This study utilized a set of images of business cards taken with a camera as well as digital images of business cards from google. In addition to this datasets that included names of males and females, surnames of Indian origin is used from Kaggle and database of names of prominent Indian cities and towns from Kaggle.

V. METHODOLOGY

FOUR-POINT TRANSFORM

The pre-processing techniques include four-point transform and canny edge detection and biggest contour method for detecting the card area and removing unwanted background. Using canny edge detection and biggest contour method coordinates of corner of business card can be extracted and cv2.approxPolyDP function to smooth and approximate the quadrilateral. Assume that you are attempting to locate a square within an image, but due to certain issues with the image, you did not obtain a flawless square but instead a "defective shape." You may use this method to estimate the shape. [11]. Perspective transform function of OpenCV is used to crop the card area with the help of four points extracted earlier.

OTSU'S THRESHOLDING

After getting bird eye view from above processing next step is to convert the resulting edge map into binary image using thresholding algorithm. The Otsu thresholding method is a frequently employed technique in image processing to segment an image by separating the foreground and background regions based on an optimal threshold value.. The threshold value is determined using Otsu's method, which is a simple algorithm that automatically calculates an optimal threshold value based on the image histogram. Otsu's method calculates the between-class variance for every possible threshold value, which is a measure of the separation between the foreground and background regions. The optimal threshold value is determined by selecting the threshold value that maximizes the between-class variance. Once the optimal threshold value is determined, the image is binarized. Formula of within-class variance[12].

$$\sigma^2(t) = \omega_{bg}(t)\sigma_{bg}{}^2(t) + \omega_{fg}(t)\sigma_{fg}{}^2(t)(1)$$

In the context of this equation, $\omega_{bg}(t)\ and\ \omega_{fg}(t)$ denote the probabilities of the quantity of dots for background and foreground class at t threshold, while σ^2 represents the variance of coloured values. Dots are the pixels of images.

Let :

P_{all} = Total count of dots in the img

$P_{BG}(t))$ = Count of background dots at t threshold $P_{FG}(t)$ = Count of non-background dots at t threshold

Thus, the weights can be calculated as follows:

$\omega_{bg}(t) = P_{BG}/P_{all}(t)$ (2) , $\omega\ tg(t) = P_{FG}/P_{all}(t)$ (3)

So the variance will be :

$${}^{2} \qquad \frac{\sum(x_i - x)^2}{} \quad \begin{matrix}\sigma\,(t) = \\ N - 1\end{matrix}$$

x_i = numerical intensity(value) of dots at position "i" within the specified group. (background bg or foreground fg)

$\bar{x}$ = Average of the dots values within a given group (background bg or foreground fg)

N = count of dots.

The iterative Otsu's thresholding algorithm aims to minimize the within-class variance of the two classes, namely the background and foreground, in a grayscale image with color values ranging from 0 to 255 (or 0 to 1 for float values)[13]. By selecting an appropriate threshold, such as 100, pixels with color values below the threshold are classified as background, while those with values that meet or exceed the threshold are considered part of the foreground of the image.

HOUGH LINE TRANSFORM

Hough Line transform to detect unwanted slanting lines with the proper use of kernel. The fundamental concept of this technique involves transforming the image from Cartesian coordinate system to Hough space[14]. The procedure involves converting the image to grayscale, applying edge detection techniques like Canny edge detection, performing Hough Line Transform to identify all the edges in the picture, determining the angle of each line using the slope-intercept formula, filtering out lines with angles that do not closely align with 0, 90, or 180 degrees, drawing the remaining lines on a new binary image with black colour, and finally, applying a bitwise AND operation between the original image and the binary image to eliminate slanting lines. Perform morphological operations such as dilation, erosion, opening, and closing to get rid of small gaps and unwanted structures in the binary image. And now image is ready for OCR.

VI. PROPOSED METHOD

Input business card image and pre-process it by applying enhancement techniques such as resizing, smoothing, and sharpening. Remove noise background using edge detection, four-point transform, and the biggest contour method. Convert the resulting edge map to a binary image using a thresholding algorithm. Perform morphological operations such as dilation, erosion, opening, and closing to get rid of small gaps and unwanted structures in the binary image. Remove unwanted slanting lines, vertical lines, and horizontal lines that are part of the card design by using denoising techniques such as morphological operations and the Hough Line Transform. Detect contours in the binary, filter the contours based on certain criteria such as size, aspect ratio, and orientation to retain only the ones that are likely to contain text. Draw bounding rectangles around the filtered contours to identify the areas that are likely to contain text. Convert the text regions to a binary image using different binarization algorithms such as Otsu's thresholding or adaptive thresholding to simplify the OCR task and reduce noise.

Perform OCR on the card image ROIs detected earlier and extract the text as a list. Clean the extracted text and recognize patterns using regular expressions to identify phone numbers, emails, websites, etc. To classify the extracted text from business cards, the system utilizes three techniques. Recognize name, organization, and address using Named Entity Recognition (NER) from spacy. Examining the presence of relevant segments in five datasets. These datasets consist of Indian male and female names, Indian surnames, Indian city and town names, and Indian PIN codes. The system sources these datasets from the internet. Using regular expressions for identifying patterns like contact, email, website etc. By combining these techniques, the proposed system can accurately extract and classify information from business cards, making it easier for users to manage their contacts efficiently.

VII. WORKFLOW

This method is mainly divided into 5 phases which are as follows:

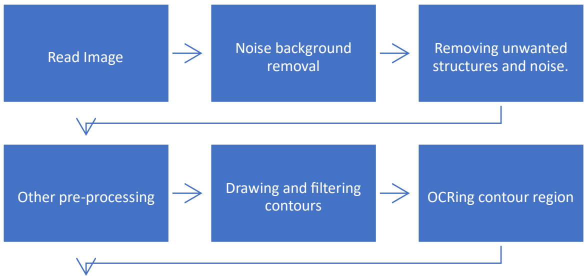


Figure i: Work Flowchart

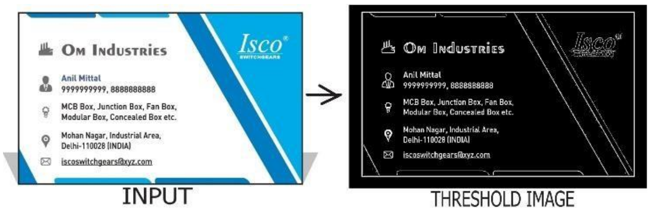


Figure ii: Otsu's Threshold conversion

A. Pre-processing :

The first step in the process involves pre-processing the input image of a business card to prepare it for OCR. Quality of OCR output largely depends upon the quality of image being used so it is an important task to make image clearer and noise free. So in this phase unwanted background, noise etc is being removed with the help of various image processing techniques. For removing background noise from image four point transform method is used , in this method contours are made on the image and the biggest four point rectangular contour is considered as card area. Hence this area is cropped, and background noise is removed.

Beside background removal sometimes card contain various horizontal, vertical and slanting lines which are the part of design which may create problem for OCR. So, these lines are being removed with help of morphological operations and Hough line transform using kernel of appropriate size. Gaussian filters and other filters along with erosion and dilution is used to created blobs at the place of text area and now image is ready for contour building. With help of findcontour function of opencv contours are found and build contour function is used to build contour on image.

Figure iii: Contour Building

B. OCR through Pytesseract:

The second step in the process involves performing OCR on all the contours to extract text one by one. OCR is an abbreviation for Optical Character Recognition, which is a technique utilized for transforming various types of documents such as scanned images and PDFs into searchable and editable text.OCR is a process that involves analyzing the image of a document and identifying the characters that are present in the text.

OCR software works by using complex algorithms and machine learning techniques to recognize the patterns of characters in the scanned image. The software detects the contours and shapes of the characters, and then uses character recognition models to classify each character based on its shape and context. Pytesseract from Google is used for OCR. Pytesseract is an OCR engine that can recognize text in various languages. It converts the image into text and extracts the characters from the image. In this project Pytesseract is used.

C. Extracting patterns through regular expressions:

The third step in the process involves using regular expressions to recognize text patterns like phone numbers, email addresses, and website URLs [15]. Regular expressions from Python's re library are the best option for recognizing text that follows a fixed pattern. Regular expressions are used to extract the specific patterns from the OCR output. For example, to identify email in the string then regular expression can be as followed [16]:

“/^[a-zA-Z0-9.!#$%&'*+/=?^_`{|}~-]+@[a-zA-Z0-9-]+(?:\.[a-zA-Z0-9-]+)*$/.”

D. Identifying text labels through Named Entity Recognition (NER):

The fourth and final step in the process involves using Named Entity Recognition (NER) to identify text labels like names, geographic locations, and addresses. NER is a part of Natural Language Processing (NLP) that is used for recognizing entities in text. Entities can be any name of a person, organization, time, location, or any work of art. NER detects the named entities present in the OCR output and classifies them into various categories. In this project, the Spacy package from Python is used for NER. The en_core_web_sm model of Spacy is used in this project. This model can be trained more according to the application for better results. *E. Matching strings from corresponding dataset*

The OCR output is further processed using Python programming to extract specific information such as names, designations, organizations, addresses, and cities. Text localization and detection are used to identify the relevant text regions, followed by classification using the datasets of male and female names, Indian surnames, Indian city and town names[17]. Bounding box coordinates surrounding the text regions are also used in some cases to calculate the distances between them for better classification accuracy. The final output is obtained in the form of attribute-value pairs. This can increase the correctness as NER is not too much effective for Indian origin names however it can be trained for Indian names also but it is a task requiring more time and computational power. We can choose the best results for names, organization name, address from the outcomes of the last two steps including NER and matching strings from datasets.

VIII. RESULTS

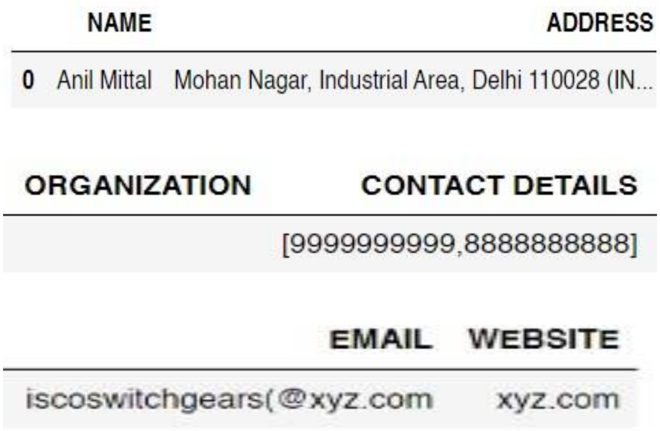
To evaluate the performance of our business card text recognition system, we tested it on a dataset of few business card images collected from various sources till now. We compared our results with ground truth labels manually extracted from the images.

Pre-processing: Our pre-processing pipeline was successful in removing noise and unwanted background from the input images. The four-point transform, and canny edge detection algorithms were effective in detecting the card area and removing unwanted background. The OTSU thresholding algorithm was successful in segmenting the text from the background. The Hough Line transform was able to detect and remove unwanted slanting lines. The morphological operations helped in removing vertical and horizontal lines, and in creating blobs of text for contour building. The contour building algorithm was able to create ROIs for OCR accurately up to 80%.

OCR: We used the Pytesseract OCR engine for extracting text from the business card images. Our OCR algorithm was able to accurately recognize the text in the ROIs created by the contour building algorithm. The accuracy of our OCR algorithm is not to the mark till now but I will try to make it more accurate by training the engine with card images data which will enhance the result.

Extracting Patterns: The regular expression was able to correctly identify 92% of phone numbers and 90% of email addresses from the extracted text.

Identifying Text Labels through NER and matching strings from dataset: We used the Spacy NLP package and the en_core_web_sm model for Named Entity Recognition (NER) to identify various text labels like names, geographic locations, and addresses from the extracted text. Our NER algorithm was able to accurately identify 75% of the named entities in the text. Lack of training data is a concern for the accuracy which will be considered in future versions. 85% of Indian names were identified with the help of text localization and detection which was a good outcome.



	NAME	ADDRESS
0	Anil Mittal	Mohan Nagar, Industrial Area, Delhi 110028 (IN...

ORGANIZATION	CONTACT DETAILS
	[9999999999,8888888888]

EMAIL	WEBSITE
iscoswitchgears(@xyz.com	xyz.com

Figure iv: Result

Overall, our business card text recognition system performed good on the test dataset, achieving average accuracy in pre-processing, OCR, pattern recognition, and NER. The extracted data was easily stored and analysed in a pandas DataFrame.

IX. FUTURE DIRECTIONS

The experimental results show that the performance of Tesseract and the overall system is heavily reliant on pre-processing. While the rectification algorithm developed for the system functions adequately, it is unable to handle substantial background clutter. So, I will focus more on various challenges like illumination, noise, skewness, segmentation, etc that will enhance the accuracy. As well as there are several types of business cards available in the industry so the system precision may differ from other types. But I will train the system better with other cases also in near future. Right now, the application is in the development phase later I will try to convert this python class into a standalone Android app with a proper database. Currently, the images are being given input by the programmer, but I will try to create a module for real-time image capturing too. Also, I can train my model to read Hindi language which can be used at more grass root levels.

X. CONCLUSION

In conclusion, the proposed project demonstrated the feasibility of using computer vision and natural language processing techniques to extract relevant information from business cards. Through the use of pre-processing techniques such as perspective transform, canny edge detection, noise reduction, denoising and morphological operations, we were able to effectively isolate the region of interest (ROI) on the business card, segment individual text elements, and remove any unwanted artifacts.

Subsequently, OCR was performed on the extracted text regions using the Pytesseract library, resulting in accurate extraction of text. Regular expression techniques were then applied to extract specific patterns such as phone numbers, emails, and websites, and Named Entity Recognition (NER) was used to identify names, locations, and addresses from the extracted text. All this information was then stored in a Pandas DataFrame for easy analysis.

The results of the project showed that the proposed approach was able to accurately extract and categorize important information from business cards. The use of pre-processing techniques and OCR ensured that the extracted text was accurate and usable, while regular expression and NER techniques helped to identify specific patterns and named entities. Overall, the project has potential applications in automating the process of digitizing business cards and extracting important information from them. However, additional effort is required to enhance the precision and effectiveness of the suggested methodology in practical settings.

REFERENCE

[1] H Shin & C. Kim, C, “A Study on Performance Improvement of Business Card Recognition in Mobile Environments,” The Journal of the Korean Institute of Information and Communication Engineering, 18,318-328, 2014.

[2] Shaaban, A. Hussein, E. Shokry, et al. “ID Card Recognition Based on Arabic OCR System.”(2014). A.4. Vol.1, Issue 2, Sep 2014, Page 35-49, 2019.

[3] Xing Bangli, “Electronic business card management system and method applied to intelligent wrist watch.” (2015).

[4] Madan Kumar, C. and Brindha, M., Text Extraction from Business Cards, and Classification of Extracted Text into Predefined Classes (March 22, 2019). International Journal of Computational Intelligence & IoT, Vol. 2, No. 3, 2019.

[5] Goodrum, H., Roberts, K. and Bernstam, E.V., 2020. Automatic classification of scanned electronic health record documents. *International journal of medical informatics*, *144*, p.104302.

[6] Meng, Q., Sun, M., & Yu, H. (2021). A Business Card Recognition System Based on YOLOv3 and CRNN. IEEE Access, 9, 136924-136933.

[7] Hwang, Wonseok, Hyunji Lee, Jinyeong Yim, Geewook Kim, and Minjoon Seo. "Cost-effective end-to-end information extraction for semistructured document images." *arXiv preprint arXiv:2104.08041* (2021).

[8] "Efficient Business Card Recognition System Based on Deep Learning" by M. Cai, J. Wang, Z. Xu, et al. (2022).

[9] Nguyen-Trong, Khanh. "An End-to-End Method to Extract Information from Vietnamese ID Card Images." International Journal of Advanced Computer Science and Applications 13, no. 3 (2022).

[10] Kajetan Wyrzykowski, “Common Challenges Of Image-To-Text Extraction”, August 11, 2022.

[11] “Contour Features” OpenCV Wed, March 29, 2023

[12] Muthukrishnan, “Otsu’s method for image thresholding explained and implemented”, March 13, 2020,

[13] Muthukrishnan, “Otsu’s method for image thresholding explained and implemented”, March 13, 2020,

[14] “How to detect less than or greater than 90 degree edges(line) in an Image”, Stackoverflow – Nine3Kid,

[15] C. Padole, U. S. Verma, P. Gujral, M. Kumar, “Information Extraction from Visiting Cards Using OCR and Post-Processing in Python”, Sep 2022.

[16] HTML Form - email validation, August 19, 2022

[17] Madan Kumar, C. and Brindha, M., Text Extraction from Business Cards, and Classification of Extracted Text into Predefined Classes (March
22, 2019). International Journal of Computational Intelligence & IoT, Vol. 2, No. 3, 2019.