International Journal of Scientific and Technical Research in Engineering (IJSTRE)
www.ijstre.com Volume 7 Issue 5 ǁ Sep-Oct 2022.
ISSN: 2581-9941

Information Extraction from Visiting Cards Using OCR and Post-Processing in Python

Chandrashekhar Padole (International Business Machines Corporation (IBM)), Uma Shankar Verma (Indraprastha Institute of Information Technology Delhi (IIITD)), Pratik Gujral (IIITD), Mrinal Kumar (IIITD), Ira Bajpai (IIITD), Diptarshi Mitra (IIITD)

Corresponding Author:
Diptarshi Mitra (e-mail: diptarshimitra@yahoo.co.in)

Abstract: This work has attempted information extraction, involving optical character recognition (OCR), from the images of twenty two visiting cards, captured with a mobile phone camera. The background of the images has been removed with the help of Python programming. In those cases where programmatic background removal either fails or produces wrong output, manual background removal has been employed. The open source Tesseract package, implemented with Python programming, has been utilized for OCR. The OCR output has been post-processed with the help of Python programming, by employing text localization and detection, followed by classification, to extract names (name1 and name2), designation, organization, address, city, PIN, country, e-mail and contact number. Here, Machine Learning/Deep Learning/Natural Language Processing techniques, which are complicated, and demand a lot of time and data, have not been used for classifying the text. Instead, for classification, the process of checking the existence of suitable parts of the text in five datasets (i.e., the datasets of the names of Indian males and females, Indian surnames, the names of Indian cities and towns, and Indian PIN codes), obtained from the Internet, has been applied, and, in suitable cases, the coordinates of the bounding box surrounding a part of the text, have also been utilized (the coordinates help in computing the distances between the bounding boxes). However, the average accuracy of the final output, yielded by this method, is quite low (34.60%).

Keywords: Visiting Card, Optical Character Recognition (OCR), Tesseract, Text Classification, Background Removal

I. Introduction

Optical character recognition (OCR) is the technique of extracting printed or handwritten text from camera images or scanned documents or image-only pdfs, and transforming the text into a machine-readable form. The process of OCR is shown in figs.- 1a and 1b.

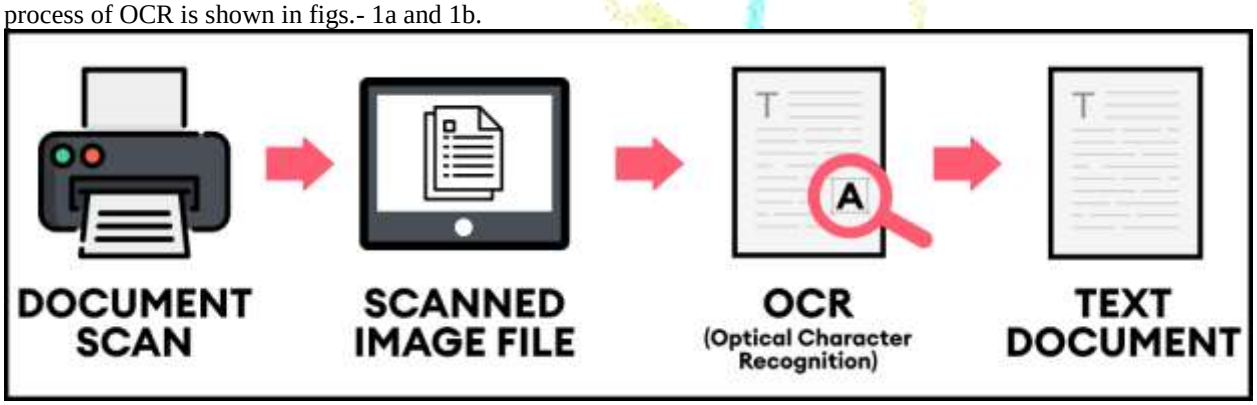

Fig.-1a: Using OCR to convert scanned data into machine-readable text
(SuperAnnotate, 2021)

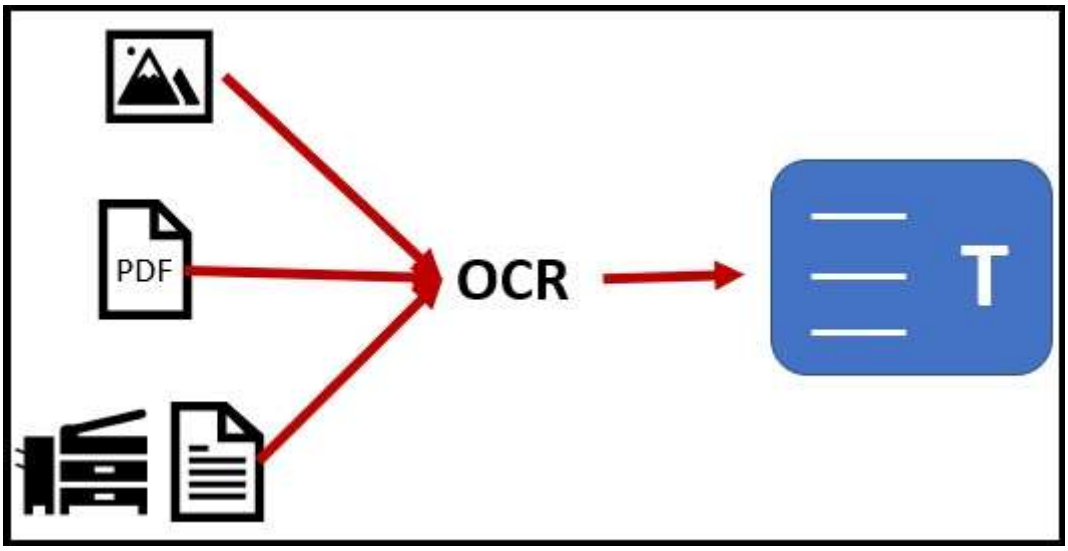

Fig.-1b: OCR accepts images, scanned documents and image-only pdfs as input
(Khandelwal, 2020)

OCR has wide applicability in various fields. As for example, OCR may be used for (i) recording information from important documents viz., bank cheques, passports, invoices, visiting cards etc., (ii) recognizing number plates of vehicles, (iii) digitizing books etc.

The objective of this work is to extract printed text from the images of visiting cards by employing OCR with the help of the Tesseract package, and subsequently classify different parts of the extracted text.

Tesseract is an open source software package used for OCR. Here, Tesseract has been employed with the help of Python programming.

Before going to the methodology and the results, it will be prudent to have a look at the work of some of the other researchers working in the field of OCR.

OCR being an important domain with applications in several fields, a considerable number of researchers worked and/or are working in this area. As for example, Kohli et al. built a model for detecting text from handwritten documents, using Neural Network and Adam Optimizer, and achieved 99.5% training accuracy and 99.0% testing accuracy (Kohli, Agarwal, & Kumar, 2022). As another instance, it may be stated that Beshah et al. developed an OCR system for both Amharic (local language of Ethiopia) and English texts, using Tesseract (Beshah & Asfawosen, 2022). Besides, Hegghammer compared the performances of Tesseract, Amazon Textract and Google Document AI, on English and Arabic texts from historical documents, and found Google Document AI to be the best among the three (Hegghammer, 2022). The review paper by Thanki et al. apprises about the work of researchers on the application of OCR in car parking control system, number plate recognition, passport recognition and information extraction etc. (Thanki, Davda, & Swaminarayan, 2021).

The application of OCR to extract information from visiting or business cards is an interesting field, and quite a number of scientists are working in this realm. For instance, Dipali et al. used Tesseract for OCR in business cards (Dipali & Lokhande, 2019). Kumar et al. also worked on business cards, applying Tesseract for OCR, and subsequently utilizing Natural Language Processing techniques, along with suitable libraries and databases, for extracting first name, last name, e-mail, organization etc. from the OCR output (Kumar & Brindha, 2019). Their results have >95% accuracies (Kumar & Brindha, 2019). Similarly, Hung et al. extracted phone number, e-mail address, name and job title from business cards, with ≥80% accuracies, by employing OCR, with the help of Tesseract, followed by Natural Language Processing methods, and appropriate databases (Hung & Linh, 2019).

This study also attempts OCR and text classification (i.e., extraction of names, designation, organization, address etc.) with visiting cards. However, though Tesseract has been employed here for OCR, Natural Language Processing techniques have not been used for classifying different parts of the text. The method adopted in this project for text classification is less time consuming and less complex.

II. Methodology

Data

In this work, the images (captured with mobile phone camera) of twenty two visiting cards belonging to Indian citizens and/or organizations, have been used.

Besides, the datasets of names of Indian males and females (URL: https://www.kaggle.com/datasets/ananysharma/indian-names-dataset), Indian surnames (URL: https://github.com/merishnaSuwal/indian_surnames_data), the names of Indian cities and towns (URL:

https://www.downloadexcelfiles.com/wo_en/download-excel-file-list-cities-towns-india#.YposlsVBzIU), and Indian PIN codes (URL: https://data.gov.in/search?title=PIN%20database), have been utilized.

Method

The method employed here is as follows:

- The images of twenty two visiting cards have been captured with a mobile phone camera.
- In the images, the background has been removed by applying the techniques of edge detection and contour processing, with the help of Python programming. While writing the code for background removal, some of the code snippets, shown in a particular website, have been used (Rosebrock, 2021).
- For some images, either the background could not be removed by programming, or the OCR output is poor after removing the background by programming. In these cases, the background has been removed manually.
- Tesseract package (along with Pytesseract wrapper) has been used to implement OCR with the help of Python programming.
- The OCR output has been post-processed, with the help of Python programming, by employing text localization and detection, followed by classification, to extract names (name1 and name2), designation, organization, address, city, PIN, country, e-mail and contact number. For classification, the technique of checking the existence of suitable parts of the text in the five aforesaid datasets (i.e., the datasets of the names of Indian males and females, Indian surnames, the names of Indian cities and towns, and Indian PIN codes), obtained from the Internet, has been employed. In addition, the coordinates of the bounding box surrounding a part of the text, have also been utilized, in suitable cases, for classification (the coordinates help in computing the distances between the bounding boxes). The result obtained after post-processing, in the form of key-value pairs, is the final output.
- The accuracy of the final output has been calculated for each card, and finally, the average accuracy has been obtained by computing the arithmetic mean of all the accuracies.

The method has been depicted diagrammatically with the help of a flowchart in fig.-2.

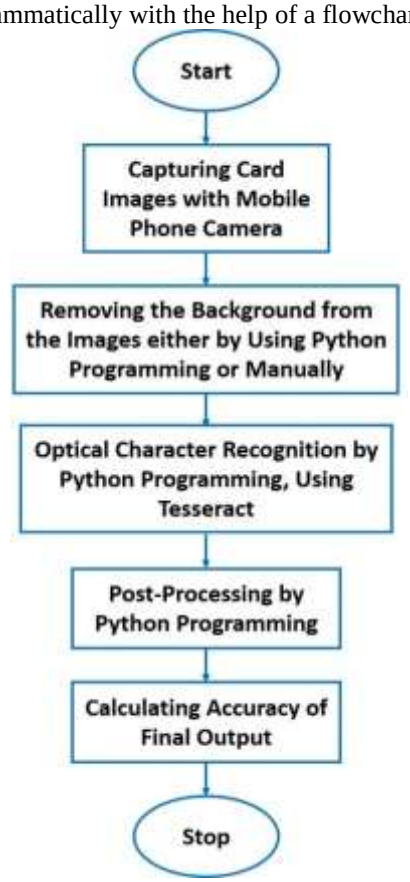

Fig.-2: Method employed in this project

Calculation of Accuracy

The formula used for calculating the accuracy (a) of the final output, is depicted in equation-1.

$$a = \frac{1}{n}\sum_{i=1}^{n}\frac{\max((cw_i - ew_i),0)}{tw_i} \quad (1)$$

where, n=number of classes (viz., name1, name2, designation etc.) for which either there is some information in the card, or some output has been generated, or both,
cw_i=number of correct words corresponding to the ith class, in the output,
ew_i=number of extra words (when the number of words in the output is more than the total number of words in the card, for a particular class) corresponding to the ith class, in the output, and
tw_i=total number of words corresponding to the ith class, in the card.

III. Results and Discussions

The images of the twenty two cards, along with the accuracies of the corresponding final outputs, are shown in fig.-3. It may be noted that the images of fifteen cards (card-1 to card-15) yielded satisfactory OCR outputs, after removing the background with the help of programming. For the images of five cards (card-16 to card-20), the background had to be removed manually, as it could not be removed with the help of programming. In case of the images of two cards (card-21 and card-22), the background had to be separated manually, as these images generated negligible outputs after removing the background with the help of programming. In fig.-3, the images of the first fifteen cards (card-1 to card-15) have been shown with background, and the images of the remaining seven cards (card-16 to card-22) have been displayed without background.

Card-1 (Accuracy: 78.70%)


Card-2 (Accuracy: 37.50%)

Card-3 (Accuracy: 40.00%)

Card-4 (Accuracy: 43.21%)

Card-5 (Accuracy: 18.81%)

Card-6 (Accuracy: 40.78%)

Card-7 (Accuracy: 36.94%)

Card-8 (Accuracy: 36.67%)


Card-9 (Accuracy: 42.60%)

Card-10 (Accuracy: 46.82%)

Card-11 (Accuracy: 33.45%)

Card-12 (Accuracy: 25.00%)

Card-13 (Accuracy: 28.61%)

Card-14 (Accuracy: 15.44%)

Card-15 (Accuracy: 1.67%)

Card-16 (Accuracy: 0.00%)

Card-17 (Accuracy: 32.62%)

Card-18 (Accuracy: 48.82%)

Card-19 (Accuracy: 28.39%)

Card-20 (Accuracy: 18.57%)

Card-21 (Accuracy: 41.51%)

Card-22 (Accuracy: 65.00%)

Fig.-3: Accuracy of the output for each card

The average accuracy has been found to be 34.60%.
Thus, on the basis of whatever has been stated and shown so far, the following inferences can be drawn:

- The technique employed for text classification (i.e., checking the existence of suitable parts of the text in the five datasets, and using the coordinates of the bounding boxes to find the distances between them, in suitable cases), could not produce a promising result.

- Among the datasets of the names of Indian males and females, Indian surnames, the names of Indian cities and towns, and Indian PIN codes, used for classification, the first four are not quite exhaustive, and hence have contributed to the poor quality of the classification output.
- Actually, the focus of this study is on finding a solution to the text classification problem, without involving complicated, data hungry and time consuming techniques of Machine Learning/Deep Learning/Natural Language Processing. Though the average accuracy, obtained here, is low (34.60%), there is scope for further improvement by using more comprehensive datasets, and including other suitable feature/s, if any, of the OCR output.

It would have been better if the accuracy of the final output, observed here, could be compared with the accuracies obtained by the other researchers who employed similar technique/s. But, though a number of studies are there, which have used Natural Language Processing techniques for post-processing (as for example, the work of Kumar et al. who observed >95% accuracies (mentioned in the *Introduction* section of the current study), or Hung et al. who noted ≥80% accuracies (also mentioned in the *Introduction* section), or Rusli et al. who got an F-score of 0.78, after using Tesseract and Pytesseract for OCR, and Natural Language Processing tools for text correction (Rusli, Adhiguna, & Irawan, 2021)), no article has been found where a procedure similar to the one used here, has been adopted. Thus, it may be presumed that the method employed in this work, has not been implemented before.

IV. Conclusion

This project has taken the images of twenty two visiting cards, captured with a mobile phone camera, as the input, and performed background removal, either programmatically or manually, followed by OCR, using Tesseract, and post-processing which includes text classification. The aim of text classification is to extract names (name1 and name2), designation, organization, address, city, PIN, country, e-mail and contact number from the OCR output.

The technique used here for text classification, which has not been found to be employed by any other researcher, and which involves checking the existence of suitable parts of the text in five datasets (datasets of the names of Indian males and females, Indian surnames, the names of Indian cities and towns, and Indian PIN codes), obtained from the Internet, and using the coordinates of the bounding boxes (a bounding box surrounds a part of the text) to find the distances between them, in suitable cases, could not generate an encouraging output (average accuracy: 34.60%). The implementation of a suitable Machine Learning/Deep Learning/Natural Language Processing method for text classification could have produced a more accurate result. However, this study intends to solve the text classification problem, without using the techniques of Machine Learning/Deep Learning/Natural Language Processing, which are complicated, and demand a lot of time and data. So, the problem of low average accuracy needs to be tackled by some other means.

One reason for the low quality of the text classification output is the fact that four among the five aforesaid datasets, employed in the classification process, are not quite exhaustive. Use of more comprehensive datasets during text classification might have produced a better output. In addition, a dataset of Indian organizations, if included in the classification process, would have improved the output further.

Besides, the coordinates of the bounding boxes of each part of the text, in the OCR output, have been used here for text classification. Inclusion of other suitable feature/s, if any, of the OCR output, may improve the accuracy.

Thus, there is scope for further improvement of the quality of the text classification output.

Moreover, it would have been better if the output could have been presented in a suitable user interface.

If possible, all these modifications will be implemented in future.

V. Acknowledgements

We are thankful to the employees of IBM for their help and support during this study.

We are also indebted to the faculty members of IIIT-Delhi for their cooperation and encouragement with regard to this work.

References

[1]. Beshah, T., & Asfawosen, A. (2022). Design and Development of an OCR System that can Convert Both Amharic and English Based Images and Scanned PDF Files into Editable Content. Journal of Management Information and Decision Sciences, 25(7S), 1–11.

[2]. Dipali, R. K., & Lokhande, D. G. (2019). Extracting Business Card Information into Contact List. International Journal of Research in Engineering, Science and Management (IJRESM), 2(6), 620–622.

[3]. Hegghammer, T. (2022). OCR with Tesseract, Amazon Textract, and Google Document AI: A Benchmarking Experiment. Journal of Computational Social Science (JCSS), 5(1), 861–882.

[4]. Hung, P. D., & Linh, D. Q. (2019). Implementing an Android Application for Automatic Vietnamese Business Card Recognition. Pattern Recognition and Image Analysis, 29(1), 156–166.
[5]. Khandelwal, R. (2020). An Introduction to Optical Character Recognition for Beginners. Retrieved from https://towardsdatascience.com/an-introduction-to-optical-character-recognition-for-beginners-14268c99d60
[6]. Kohli, H., Agarwal, J., & Kumar, M. (2022). An Improved Method for Text Detection Using Adam Optimization Algorithm. Global Transitions Proceedings, 3(1), 230–234.
[7]. Kumar, C. M., & Brindha, M. (2019). Text Extraction from Business Cards and Classification of Extracted Text into Predefined Classes. International Journal of Computational Intelligence & IoT (IJCIIoT), 2(3), 595–602.
[8]. Rosebrock, A. (2021). OCR'ing Business Cards. Retrieved from https://pyimagesearch.com/2021/11/03/ocring-business-cards/
[9]. Rusli, F. M., Adhiguna, K. A., & Irawan, H. (2021). Indonesian ID Card Extractor Using Optical Character Recognition and Natural Language Post-Processing. In 9th International Conference on Information and Communication Technology (ICoICT) (2021) (pp. 621–626). IEEE.
[10]. SuperAnnotate. (2021). What is Optical Character Recognition (OCR): Overview and Use Cases. Retrieved from https://blog.superannotate.com/ocr-overview-and-use-cases/
[11]. Thanki, J. D., Davda, P. D., & Swaminarayan, P. (2021). A Review on OCR Technology. Journal of Emerging Technologies and Innovative Research (JETIR), 8(4), 716–720.