Volume 11, Special Issue I, May 2020
International Journal of Advanced Research in Computer Science
CONFERENCE PAPER
Available Online at www.ijarcs.info
ISSN No. 0976-5697

EXTRACTING TEXT FROM THE PICTURE BY USING OCR TECHNOLOGY

K V Sheelavathy [1], Navneet Priya[2]
[1] Department of Computer Science and Engineering, REVA University, Bengaluru, India
[2] Department of Computer Science and Engineering, REVA University, Bengaluru, India
Kvsheelavathy_sheelakv@reva.edu.in, navneetpriya_navneet1617.npr@gmail.com

Abstract— Like it, all known, almost all the cell phones now existing in the world almost have all have at least good quality of back camera except to some phone and also, a few years before, technology enhances and technical industry people launched new application by using this camera on which very big documents is achieved by completely traversing this camera on it. Now, it's a strong effort to make out to upgrade this application. Now, this application can extract the text from the picture. It is a tricky assignment due to creative actions of mobile camera beside by way of hand on shaking, transform in enlightenment suitable to hand overshade, etc. the matter has extra complied through the detail that transcript alphabets which not hold diverse quality which is mainly an input necessity for panoramic icon invention. An outline is intended for some time to filter the configuration of icons depends upon the supermodel designing. This saves memory by extracting text from images by using OCR technology.

Keywords—OCR

I. INTRODUCTION

Multiple facilities in one device give some relief to the present users who have the use of various technologies for their work. One cell phone now can do various things like playing video, audio, reading books, world best dictionaries, and so many things. Now, putting a lot of effort here to make something which makes the camera even more from clicking pictures that is to make use of a camera which lowers the gap between the virtual space and physical space. This application took all images of visual codes that are written on the pages. After taking that it encodes all the text on the pages from the same image.

This application first scans the hardcopy, then by using OCR technology it extracts the text written in the hard copy. OCR software changed the whole scanned image into two types of color one is black and the other is white. After scanning the image it examines for the luminous and shadowy area, where the shadowy area is identified as characters that required to be acknowledged and luminous area are recognized as the backdrop. After that, shadows are recognized by either alphabets or numeric. It is mainly focusing on one text at a time whether it is one character or word or else block of text.
Characters identification is done by using these two main algorithms

1. Pattern recognition algorithm
2. Feature detection algorithm

This application is made by the research works with some changes which are appropriate for manuscript based matching.

1. PATTERN RECOGNITION ALGORITHM

This algorithm is used to detect the various fonts as well as formats and then that format is used to contrast and identify text whether it is in the form of alphabets and numeric in the scanned document.

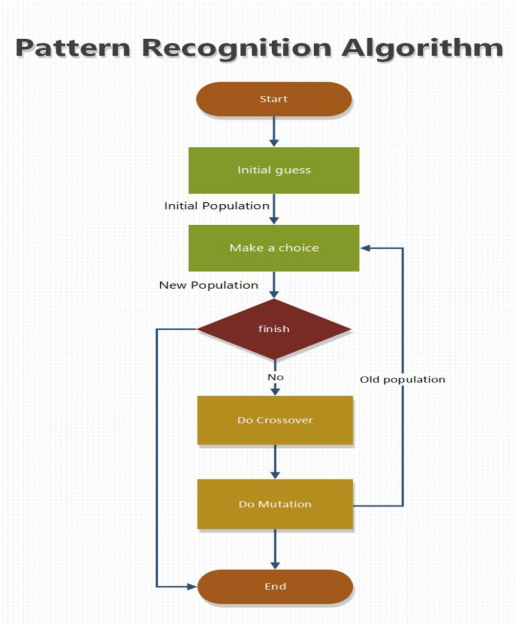

Figure 1. Pattern Recognition Algorithm

2. FEATURE DETECTION ALGORITHM

According to the OCR program some rules are there to recognize the specific features of alphabets or numeric from the scanned files. Characteristics could consist of the number of oblique lines, crossed to each other lines or upward and downward curves which are maybe in the form of the parabola in text for contrast. Let's take an example, like take a capital alphabet "H" possibly will be stored like two vertical lines and one horizontal line in between these two vertical lines. When quality is recognized, it is changed into an ASCII code which is worn by the computer's system to manage the manipulations ahead. The users and consumers should identify and correct all the basic mistakes analyzed and be sure that complicated designs were managed appropriately before saving the file for further use.

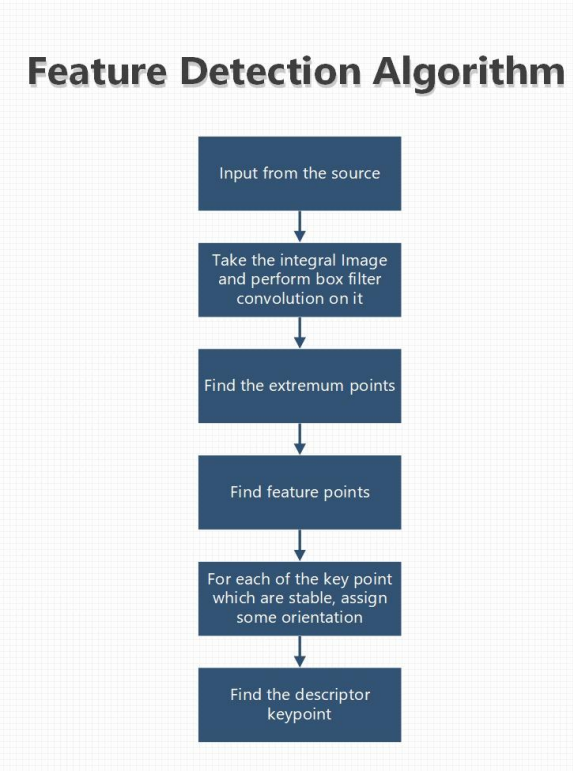

Figure 2. Feature detection Algorithm

II. LITERATURE SURVEY

- The strategy of a programmed linear textbook scanning devices, with few traits, is suggested in [1].
- Prosperous bodily design investigation is a key feature in the presentation of text identifiers and a large number of other different applications. To benefit and calculate the act of new Arabic physical layout exploration answers, [2] grants the outcomes of the competition as well as implementation.
- Experiments in [3] signpost that for the situation of few restricted exercise data accessibility recovers the scanning and illumination accurateness of printer.
- Experiments in [4] point out that outcomes on genuine records of that paper demonstrate that indexing arrangement is in effect as well as well-organized in index age group and level.
- The huge figure of try-outs in [6] illustrates that the future retrieval technique in [6] is modest and viable. That skill not only saves the unique kind of these ancient archives, it can also help persons to get these files visibly and professionally.
- An original technique to precise view and symmetrical alterations as fine as to section and polish sheet edges, which is free of file matters and also it does not need any extra hardware or numerous pictures is conversed in [5].

III. FIGURES AND TABLES

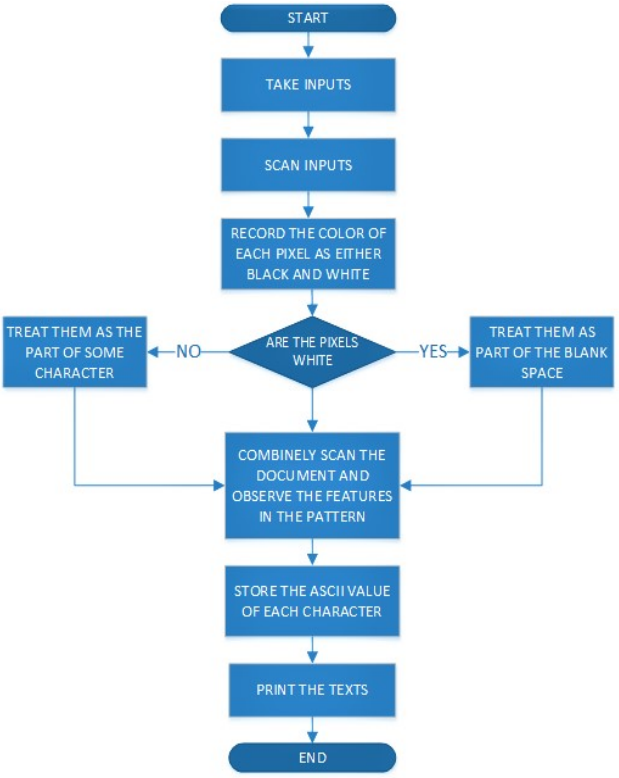

Figure 3. Inner Design and Working Principle of the proposed scanning application

IV. Abbreviations and Acronyms

OCR – Optical character recognition
ASCII–Differencing American standard code for information interchange

V. CONCLUSION

To save the money size, it's just an effort to make the application which not only able to scan the documents but also be able to extract the text from those documents. This saves the memory and also paper wastage because with the help of this application no need to write something virtually because this application scanned evens the virtual copy directly from the notepad. The foremost focus of this technical project is to mechanize the facts which are scanning and made that app capable so that this function is accomplished of taking and accepting the necessary information by that app itself and will save it. Its upcoming possibility is to keep the bill data such that it is online access to formulate the data obtainable across different devices. Another future scope of this application is that till now, in all the scanning applications, they scan images only and they won't be able to edit once images get scanned but in this application, there is no need for paper. Once scanned then the bill is saved locally and shows all those items in a table together.

VI. FUTURE SCOPE

The future scope of this application is to save the contents locally in the mobile and also save its content online by using Google's cloud vision and also make the application to make the user able to access the contents from anywhere online if they get lost their mobile or in any other cases.

Another future scope is like google drive here in this application anyone put their contents and upload that in this device and be able to share the links of any contents. No need to share the whole contents.

The last scope is to shows all the contents that it's user keeps in it and save in it in the form of a table.

VII. ACKNOWLEDGMENT

I scope my honor towards my college REVA University for providing me the chance to move forward with this new project. I am thankful to my department professors who helped me a lot through my project. I scope my wholehearted honor towards my guide Prof. Sheelavathy for guiding me by giving me all the technical ideas necessary to make this project. At last, I am thankful to all who have supported functioning affectionately in the implementation of this project.

VIII. REFERENCES

[1] Osama W. Ata, Ahmad Dadou, Mohammad Ismail, "Automated Linear Book-Scanning System (ALBS)", 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT

[2] Randa Elanwar, Margrit Betke, "Physical Layout Analysis of Scanned Arabic Books", 2018 IEEE 2nd International Workshop on ASAR

[3] Sharad Joshi, Mohit Lamba, Vivek Goyal, Nitin Khanna, "Augmented data and improved noise residual-based CNN for printer source identification", 2018 IEEE

[4] Ragavan N, "Efficient Key Hash Indexing Scheme with Page Rank for Category Based Search Engine Big Data", 2017 IEEE International Conference on Intelligent Techniques in control, Optimization and Signal Processing

[5] Neng Quan, Xiaolong Zhou, Shengyong Chen, "Scan Paperback books by a Camera", International Conference on Information and Automation, August 2016

[6] Yan Jiang, Yingjie Xia, Zhiquan Feng, Jinping Li, "Research on the Recovery Technology of Scanned Image of Obsolete Document", 2017 4th International Conference on ICCSS

IX. AUTHOR PROFILE

I am Navneet Priya pursuing my B.Tech from REVA University. Currently, I am in the 6th semester and I am writing this project in the IJEAT format.