OCR With Background Image Elimination-A Survey

Damini J. Patel
P. G. scholar
CSE Department
Gujarat Technological University,
Ahmedabad, India

Prof. Shital V. Patel
Professor
CSE Department
Gujarat Technological University,
Ahmedabad, India

ABSTRACT

Optical Character Recognition (OCR) system converts scanned input documents into an Editable Text Document. This report presents for OCR with background image elimination, the various stages (techniques) in OCR, Accuracy of it and use of different Software's to implement this technique. It is widely used as a form for food package, Invoices, FMCG, on metal parts, passport documents, printout of static data etc. An OCR system enables us to feed a book or a magazine article directly into an Electronic Computer File and edit it. The various stages of an OCR are: Upload a Scanned Image from the Computer, Segmentation Process in which we extract the text zone from the image, Recognition of the text and the last which is Post Processing Process in which the output of the previous stage goes through the Error Detection and Correction Phase. In this project we use the method is effective in removing the background of image and enhance the performance of OCR. The output image is clean after the background elimination.

Keywords—*OCR, Background image, image pre-processing, OpenCV*

1. INTRODUCTION

OCR stands for Optical Character Recognition. A person is able to see images because of the communication between our eyes and brain. Our eyes act as an optical mechanism and the images seen by our eyes are an input for our brain and the ability to understand visualize these images varies from person to person. Similarly we have the technology known as OCR, where OCR stands for Optical Character Recognition, which through its automated mechanism allows easier recognition of character and it's processing. [1]

Earlier scanners were the only working OCR application available in the market. The main disadvantage of scanners was that it was not portable and it takes a lot of time to capture an image. [2]

But with today's devices having better processing speeds, larger internal memory and an excellent back camera, researchers have dared to think of running OCR applications on devices such as smart phones for having real time imaging results. Applications such as Cam Scanner and Google translate are the prime examples of Optical character Recognition application. It also showcases the fact that this OCR technology can be put to use in a wide array of streams and hence is a very important concept which requires more attention towards research. [3]

HOW OUR OCR WORKS

a. *What is OCR*

OCR allows for automatically recognizing characters through an optical mechanism. It is capable of recognizing both handwritten and printed text. Its performance can be judged based on the quality of the documents and the camera being used to capture the raw image. OCR system is so designed that it processes images with contain more text with very less number of graphic element. [4]

As mentioned before, most of the character recognition programs and algorithms will be working efficiently only on the images which are captured using a scanner or a digital camera and run on a computer software. But since the size and portability were the factors which were hampering further growth and usability of this technology, in order to overcome the above mentioned limitations, a character recognition system based on android devices is proposed. [5]

OCR as a technology that enables us to convert various types of documents such as scanned papers, PDF files or images captured by a digital camera into editable and searchable data. A point worth noting is that the images captured by a digital camera differ from scanned documents or images as they often have distortions in their captured images. These distortions and noise makes it difficult to recognize the text accurately. Pre-processing is done on the image to improve the accuracy of text recognition. [6]

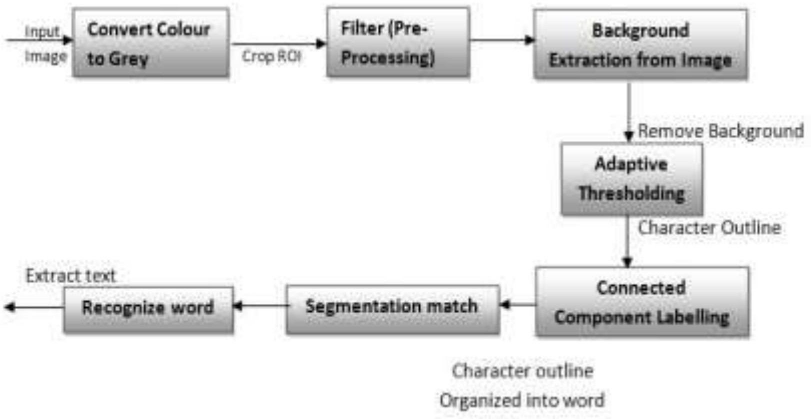
b. *Our OCR Works*

In OCR Many document images are embedded with background images. it causes difficulties for OCR applications. Some parts of the background image could be bound-boxed as characters, which leads to immediate wrong recognition and causes troubles in the following processing steps of the OCR pipeline.

Therefore, I proposed a new method for background image elimination.it is very important to pre-process the documents by removing the background images before text detection. The existing work is considers only good quality of printed document is considered without any touching or broken characters.

The existing method is effective in removing the background image and thus enhances the performance of OCR. This method is based on the difference of the values of the R, G, B colors in background image pixels. The experiments showed that the output image is clean after using the preprocessing.

OpenCV is chosen for OCR because of widespread, approbation, extensibility, and fexibilit. OpenCV is an open source library for Image Processing. It is available on many operating system. It is most accurate open sourse library.[7]



Figuer 1 Our OCR Work With OpenCV

In this convert the input image into binary format using adaptive thesholding outlines of copmonents are stored on connected component analysis.Nesting of outlines is done which gathers the outlines together to form a Blob.Text lines are analyzed for fixed pitch and proportional text.

Then the lines are broken into words by analysis according to the character spacing. Fixed pitch ischopped in character cells and proportional text is broken into words by definite spaces and fuzzy spaces.

In this recognises a word in two passes, that is,it tries to recognize the words in the first pass. If the match is found, then the found word is passed on to the Adaptive Classifier, which recognizes the text more accurately.

During the second pass, the words which were not at all recognized or were not well recognised in the first pass are recognized again through a run over through the page. Finally resolves fuzzy spaces. To locate small and capital text, checks alternative hypothesis for x-height. [3]

OCR technology has a broad range of applications in document processing. Many document images are embedded with background images, e.g., checks, deposit books, drive licenses, passports, certificates, etc. While the background image enhances the document's security or visual effects, it causes difficulties for OCR applications. Some parts of the background image could be bound-boxed as characters, which leads to immediate wrong recognition and causes troubles in the following processing steps of the OCR pipeline. Therefore, it is very important to preprocess the documents by removing the background images before text detection.

2. METHODOLOGY

The Fig.2 given below is the overallfunctioning of Optical Character Recognition (OCR). The input image can be any document, live text, journals, magazines etc. The functioning of OCR contains the following steps: scanning, segmentation, pre-processing, feature extraction,recognition[5]. The input is first scanned using an Android mobile camera. This is done to digitize the document. Segmentation extracts any symbols in the text region. Noise is removed by pre-processing each symbol,and the characteristics of each symbol is extracted using feature extraction to finally recognise the text.

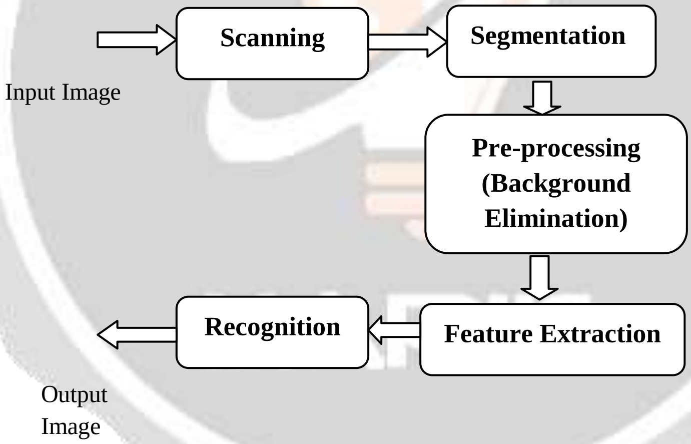


Figure 2 Overall functioning of OCR

2.1 Scanning

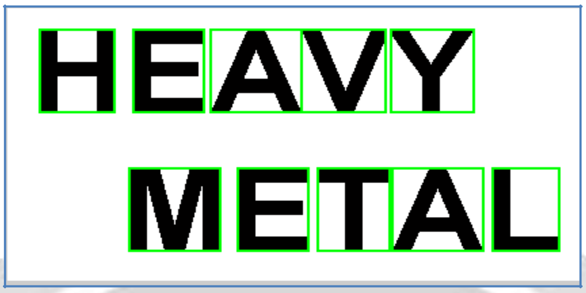
Android mobile camera is used to capture the image of document. This process is called scanning.

This is nothing but the process of scanning which converts the document into digital image. The digital image is then converted into a grayscale image using Thresholding function.

Thresholding is the process which converts multi level image into bi-level image i.e. black and white image.Black is represented if the gray level is below the threshold level, and it is represented by white if the gray level is above the threshold level. This makes it easier to detect the text regions in an image. It also saves a lot of memory space and processing time.[10]

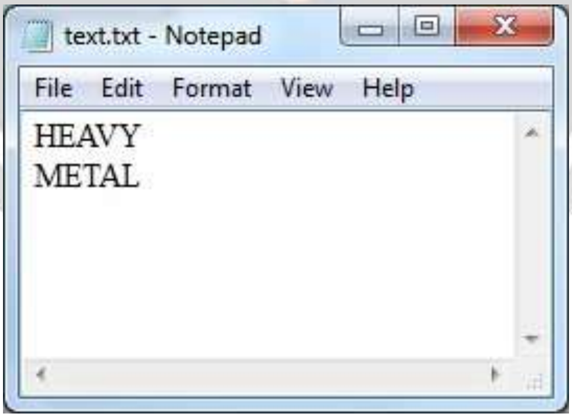
2.2 Segmentation

Regions of text is detected using the process of segmentation. It differentiates the text from other graphical elements in the document. Splits and joints can cause confusion between text and graphic elements in the document resulting in incorrect segmentation of the text.[5] This generally occurs due to poor scanning which increases the noise in the digital document. Joints in characters occurs when the document is scanned at low threshold and splits occurs when the document is scanned at high threshold.



Figuer3. Example of charecter segmentation[9]

Figuer4. Example of charecter Recognition [9]



HEAVY
METAL

Figuer5. Result of applying OCR [9]

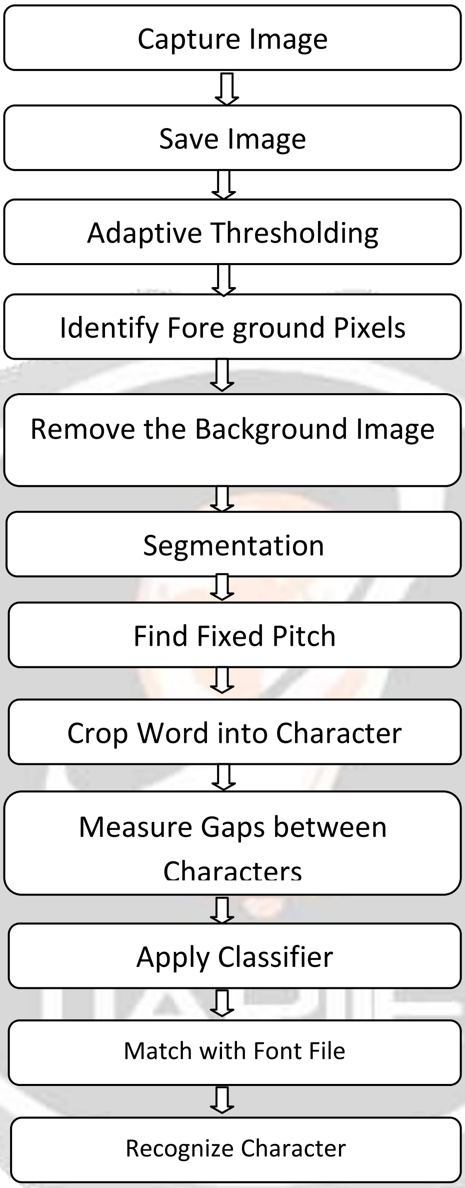
2.3 Pre-Processing

During scanning stage, some noise is produced in the scanned image. This results in poor recognition of characters. This noise can be reduced by pre-processing.Pre-processing is done using smoothing and normalization.Smoothing is done on the image using filling and thinning techniques. Normalization is responsibleto handle uniform size, slant and skew correction. [5]

2.4 Feature Extraction

Feature extraction refers to the extraction of features of symbols from the image. In this step, only important attributes are taken into account and any unnecessary attributes are ignored. This technique takes into account the abstract features present in the character. Spaces, lines, intersections etc are some of the abstract features.

Flow Chart For Our OCR System



Figuer.6 Flow Chart For Our OCR System

Feature extraction is done using OpenCV algorithm. OpenCV algorithm is used to implement feature extraction.[6]

2.5 Recognition

OCR system uses OpenCV to identify characters from the image foreground pixels also called as blobs and recognizes the lines. These lines are then recognized into words or characters.In this phase the image is converted into character stream which represents letters. [7]

3. APPLICATIONS OF OCR

- Data Entry and Text Entry
- Process Automation[14]
- Banking – Read and transfer correct amount of money from printed cheques.[15]
- Food package
- FMCG
- Automatic number plate recognition[16]
- Legal – digitize paper documents[8]

4. CONCLUSION

The presented work is effective in removing background image to improve performance of OCR using adaptive threshold. This method is use pre-processing and contour tracking algorithms. IN the pre-processing we have use a Gaussian blur and Goble threshold and find contour method and remove the background of image.It improve charecter qulity.If charecter are cut then do join method.

5. REFERENCES

[1] Sravan Ch, ShivankuMahna , NirbhayKashyap," Optical Character Recognition on Handheld Devices",International Journal of Computer Applications (0975 – 8887).

[2]Ali Farhat*, Ali Al-Zawqari, Abdulhadi Al-Qahtani, Omar Hommos, Faycal Bensaali and Abbes Amira, "OCR Based Feature Extraction and Template Matching Algorithms for Qatari Number Plate", 978-1-4673-8743-9/16/$31.00 ©2016 IEEE.

[3] Pooja Sharma, Shanu Sharma,"An analysis of Vision Based Techniques for Quality Assessment and Enhancement of Camera Captured Document Images", 978-1-4673-8203-8/16/$31.00 c 2016 IEEE.

[4] Dave Desrochers, Zhihua &U: and Apiwat Saengdeejing, "OCR Readability Study and Algorithms for Testing Partially Damaged Characters", Proceedings of 2001 lnternational Symposium on intelligent Multimedia, Wdeo and Speech Processing May 2-4 2001 Hang Kong.

[5] Heuristic-Based OCR Post-Correction for Smart Phone Applications, the University of North Carolina at Chapel Hill department of computer science honors thesis Author: Wing-Soon Wilson Lian 2009.

[6] R. Smith. "An overview of the Tesseract OCR Engine." Proc 9th Int.Conf. on Document Analysis and Recognition, IEEE, Curitiba, Brazil,Sep 2007

[7] The Tesseract open source OCR engine, http://code.google.com/p/tesseract-ocr. International Journal of Computer Applications (0975 – 8887) Volume 115 – No. 22, April 2015 13

[8]R. Smith. "An overview of the Tesseract OCR Engine." Proc 9th Int.Conf. on Document Analysis and Recognition, IEEE, Curitiba, Brazil,Sep 2007

[9] "α-Soft: An English Language OCR", 2010 Second InternationalConference on Computer Engineering and Applications. Junaid Tariq,Umar Nauman Muhammad UmairNaru.

[10] A survey of modern optical character recognition techniques (DRAFT), February 2004

[11] Mrs. B.Vani, Ms. M. Shyni Beaulah, "High accuracy Optical Character Recognition algorithms using learning array of ANN‖, 2014 International Conference on Circuit, Power and Computing Technologies [ICCPCT].

[12]Norizam Sulaiman,Sri Nor Hafidah Mohammad Jalani, Mahfuzah Mustafa, Kamarul Hawari, “Development of Automatic Vehicle Plate Detection System”, 2013 IEEE 3rd International Conference on System Engineering and Technology, 19 - 20 Aug. 2013.

[13] Rohollah Mazrae Khoshki, Subramaniam Ganesan, “Improved Automatic License Plate Recognition (ALPR) system based on single pass Connected Component Labeling (CCL)and reign property function”, 978-1-4799-8802-0/15/$31.00 ©2015 IEEE.

[14]Archana S.Sawant, Prof. D.G.Chougule, “Script Independent Text Pre-processing and Segmentation for OCR ”, International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO) – 2015.

[15]Ayatullah Faruk Mollah, Nabamita Majumder, Subhadip Basu, Mita Nasipuri4, “Design of an Optical Character Recognition System for Camerabased Handheld Devices”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 4, No 1, July 2011.

[16]Norizam Sulaiman,Sri Nor Hafidah Mohammad Jalani, Mahfuzah Mustafa, Kamarul Hawari, “Development of Automatic Vehicle Plate Detection System”, 2013 IEEE 3rd International Conference on System Engineering and Technology, 19 - 20 Aug. 2013.