A Project Report on

OCR Model for Handwritten Image

by

Anuja Velaskar(16102042)
Nidhi Munavalli(16102049)
Apurva Waingaonkar(16102050)

Under the Guidance of

Sofiya Mujawar

Department of Computer Engineering
A.P. Shah Institute of Technology
G.B.Road,Kasarvadavli, Thane(W), Mumbai-400615
UNIVERSITY OF MUMBAI

Academic Year 2018-2019

Approval

This Project Report entitled **"OCR Model for handwritten image"** Submitted by **"Anuja Velaskar"(16102042),"Apurva Waingankar"(16102050),"Nidhi Munavalli"(16102049)",** is approved for the partial fulfillment of the requirement for the award of the degree of **Bachelor of Engineering** in **Computer Engineering** from **University of Mumbai**.

Sofiya Mujawar
Guide

Prof. Sachin Malave
Head Department of Computer Engineering

Place:A.P.Shah Institute of Technology, Thane
Date: 11 April 2019

CERTIFICATE

This is to certify that the project entitled **"OCR Model for handwritten image"** submitted by **"Anuja Velaskar"(16102042) ,"Apurva Waingankar" (16102050),"Nidhi Munavalli" (16102049),"** for the **Computer Engineering.** To the University of Mumbai, is a bonafide work carried out during academic year 2017-2018.

Sofiya Mujawar
Guide

Prof. Sachin Malave
Head Department of Computer Engineering

Dr. Uttam D.Kolekar
Principal

External Examiner(s)

1.

2.

Place:A.P.Shah Institute of Technology, Thane
Date:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, We have adequately cited and referenced the original sources. We also declare that We have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

———————————————

(Signature)

———————————————

Anuja Velaskar(16102042)
Apurva Waingankar (16102050)
Nidhi Munavalli (16102049)

Date

Abstract

There are many cultural, governmental, commercial and educational organisation that manage large number of manuscript textual information. Hindi being the national language such organisation include Hindi hand written documents.

Text line segmentation in official handwritten documents remains an open document analysis problem. Handwritten documents are scanned and the text-line is segmented according to the object of interest. In segmentation stage, the input undergoes pre-processing steps and connected components are found.

These connected components undergo bottom up grouping for the purpose of energy minimization. This is an input for optical character recognition.

Contents

List of Figures

List of Abbreviations

OCR: Optical character recognition
CNN: Convolutional Neural Network
OpenCV :Open Source Computer Vision

Chapter 1

Introduction

OCR, or optical character recognition, is one of the earliest addressed computer vision tasks, since in some aspects it does not require deep learning. Therefore there were different OCR implementations even before the deep learning boom in 2012. This makes many people think the OCR challenge is “solved”, it is no longer challenging. Another belief which comes from similar sources is that OCR does not require deep learning, or in other words, using deep learning for OCR is an overkill.

Text line segmentation of a document image is considered as a critical stage towards unconstrained handwritten document recognition. Line segmentation is the first and the most critical pre-processing step for a document recognition, followed by word segmentation, word recognition and other indexing steps. Different types of handritten documents give arise to different types of problem.These problems might occur due to different writing styles of different people, different scripts of languages,overlapping of words,adjacent line touching,etc. In this paper we concentrate only on the text

The following are the steps of OCR model

A) Input Image.

B) Segmentation.

C) Background Cleaning.

D) Skew Correction.

The above steps gives the overview of the proposed system, a handwritten text document undergoes segmentation whose output is given to the background cleaning. In this stage, all the noise is removed. Then the skew is detected and corrected in skew correction stage.

Chapter 2

Literature Review

SN	PAPER TITLE	AUTHOR	SUMMARY
1	Kannada text line extraction minimization and skew correction (2014)	1.Sunanda dixit 2.Suresh Narayan 3.Mahesh bellur	1.Segmentation of handwritten Document. 2. Extraction of handwritten documents. 3.This work also uses skew correction of the extracted text line.
2	Line and Ligature Segmentation of Urdu Nastaleeq Text (2017)	1.IBRAR AHMAD1, 2.XIAOJIE WANG1, RUIFAN LI1, MANZOOR AHMED2, AND RAHAT ULLAH3	1.The proposal mainly introduces two algorithms for line and ligature segmentation of Nastaleeq text images. 2. The proposed line segmentation algorithm places dots and diacritics more accurately as compared to Prevailing work relied more on zonal information and heuristics for line and ligature segmentation, respectively.

SN	PAPER TITLE	AUTHOR	SUMMARY
3.	Text line segmentation of hand written document of hindi and english (2014)	1.Sunanda dixit 2.Sneha 3.Nilotpal utkalit 4.Suresh h n	1.In this paper a method to detect and segment unconstrained hand written document written in English and hind where document image is first binarized and connected component are identified 2.Based on hough lines the text line are identified. 3.Skew angle is determined and than the skewness is minimized segmentation is then performed and the result is than refined by removing the noise which basically comprises component from adjacent line

Chapter 3

IMAGE PREPROCESSING STEPS

The following are the preprocessing steps performed before segmentation of the image:

An input image is taken, for image preprocessing-resize it accordingly and then increase or decrease the contrast and brightness of the image as per required.
For more accuracy, the image is converted into a grayscale.
Next binarisation is done, here binary image is created which will be helpful in segmentation processes.
Two basic morphological operations are used: Dilation and Erosion
Dilation: It increases the object area and is used to accentuate features.
Erosion: Erodes away the boundaries of foreground objects and is used to diminish the feature of an image.
Noise Removal: Noise is generally considered to be a random variable with zero means. This step clears the noise in the image and Increases the accuracy.

LEVELS OF TEXT SEGMENTATION
Text image segmentation can be achieved at three levels. As we move at different levels of text segmentation hierarchy, we obtain specifically finer details. Using all the three levels is not compulsory. Segmentation at any of these levels directly depends on the nature of the application. More the details required for the image, the more is the level of segmentation.

LINE SEGMENTATION

Line segmentation is the first and a primilary step for text based image segmentation. It includes horizontal scanning of the image, pixel-row by pixel-row from left to right and top to bottom. At each pixel the intensity is tested. Depending on the values of the pixels we group pixels into multiple regions from the entire image. The different region indicates different content in the image file.

WORD SEGMENTATION

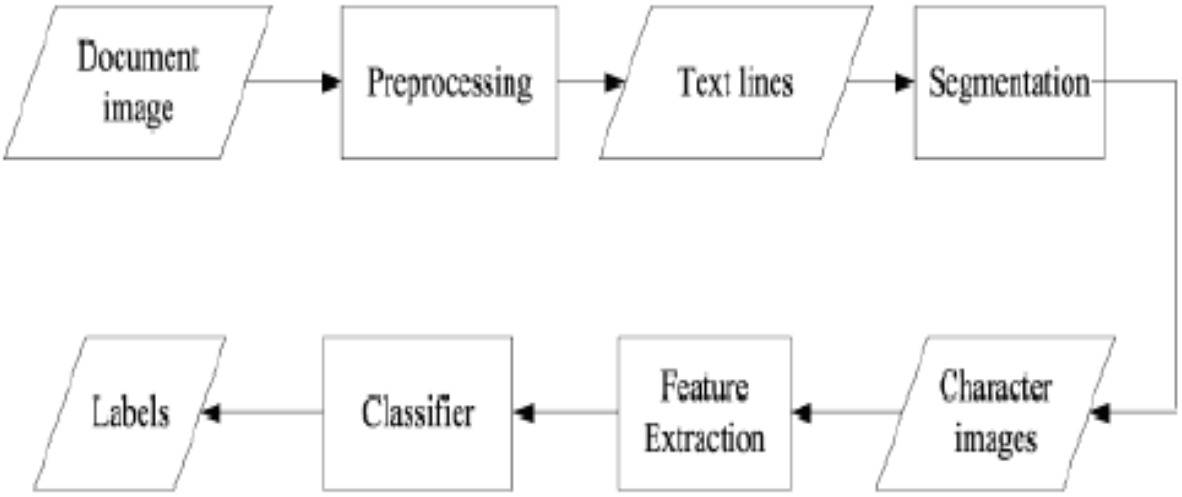
Word segmentation is the next level of segmentation. It includes vertical scanning of the image, pixel-row by pixel-row from left to right and top to bottom. At each pixel the intensity is tested. Depending on the values of the pixels we group pixels into multiple regions from the entire image. The different region indicates different content in the image file. Subsequently the desired content can be extracted.

CHARACTER SEGMENTATION

Character segmentation is the final level for text based image segmentation. It is similar to in operations as word segmentation. A few precautions should be followed while performing character segmentation. Figure 2 shows one such problem. The segments are not accurate, as “h” is extracted as “l” and “i”. Such errors are undesirable. Another precaution is of ligatures. If the text image contains a cursive type font then while segmenting the ligature should be separated for better efficiency

Chapter 4

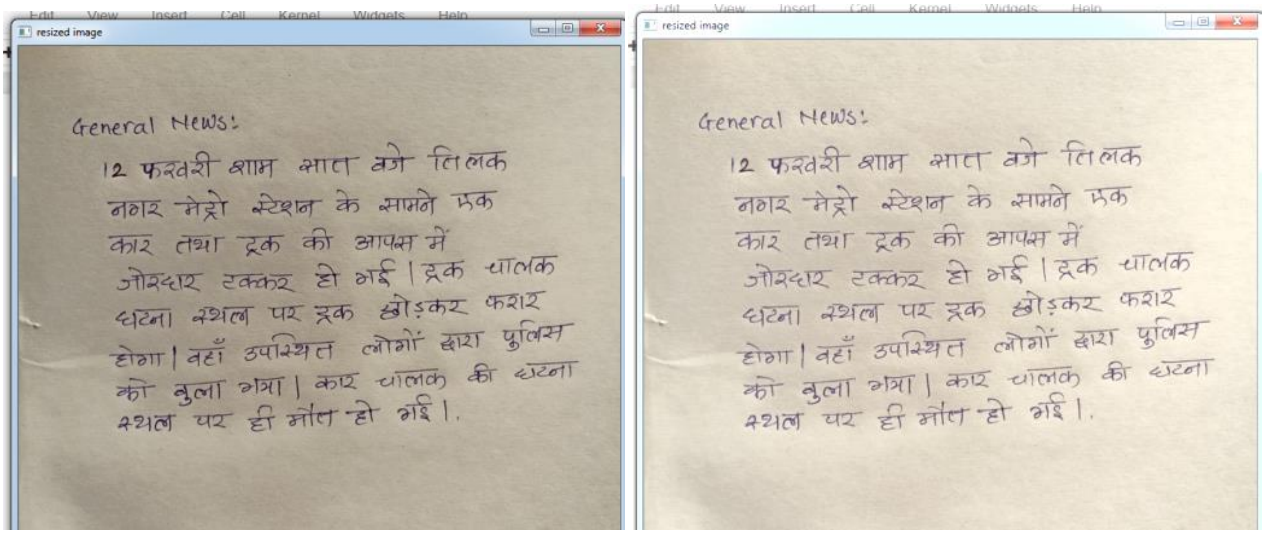
Flow chart of OCR model:-



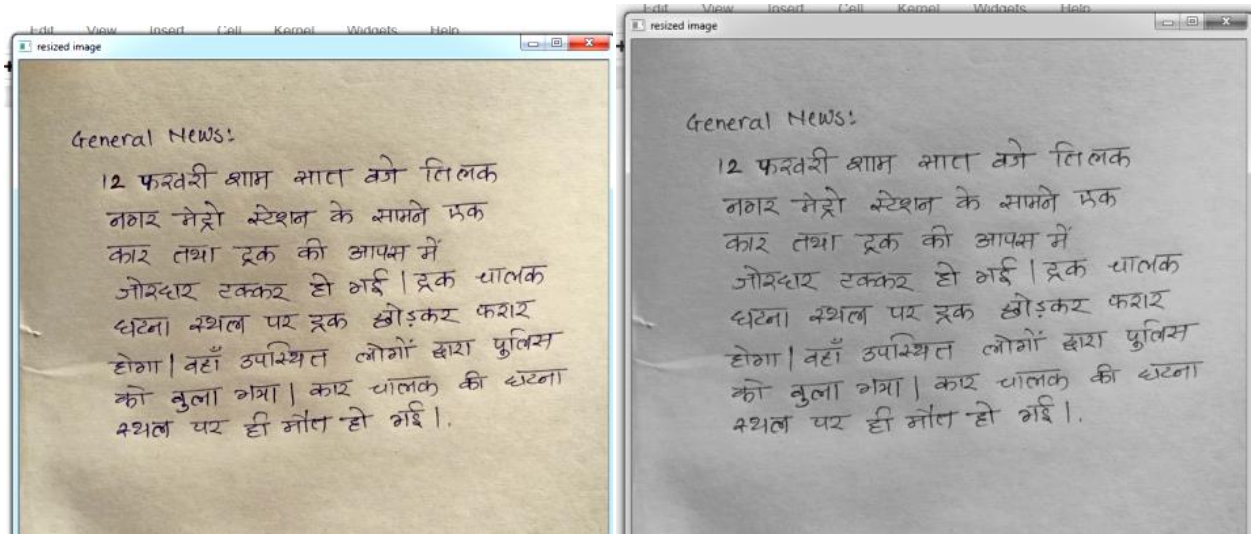
Result

For the purpose of evaluation, lines are considered to be detected when it is successfully able to detect the punctuation like characters and also components from the adjacent lines are successfully separated, or is minimal in case the lines are very close to each other.

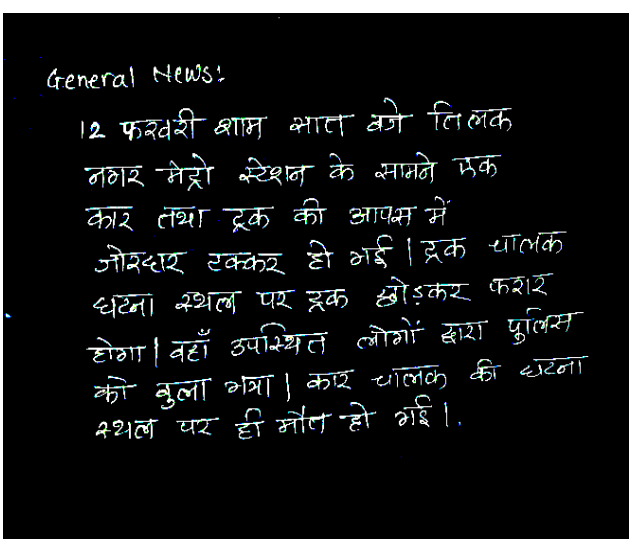
Experiment is conducted on a scanned handwritten document. The document undergoes binarisation as a part of preprocessing. Then the noise is removed based on the area for the accurate detection of text lines. The text lines are detected and segmented; is detected line is indicated by a bounding box.

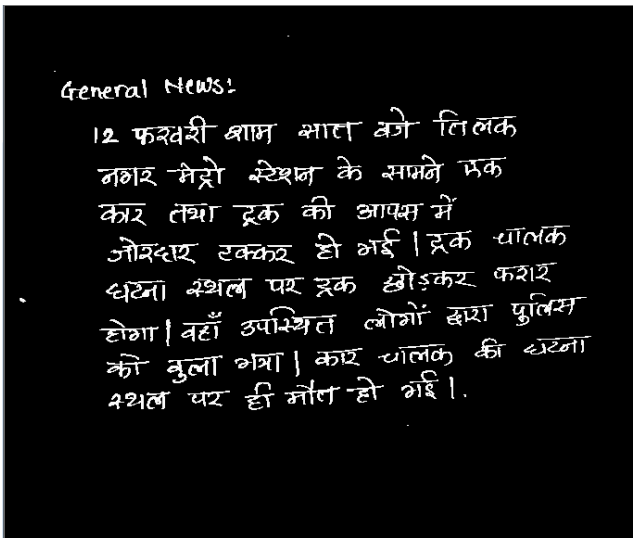


Preprocessing step: images of resizing and brightness

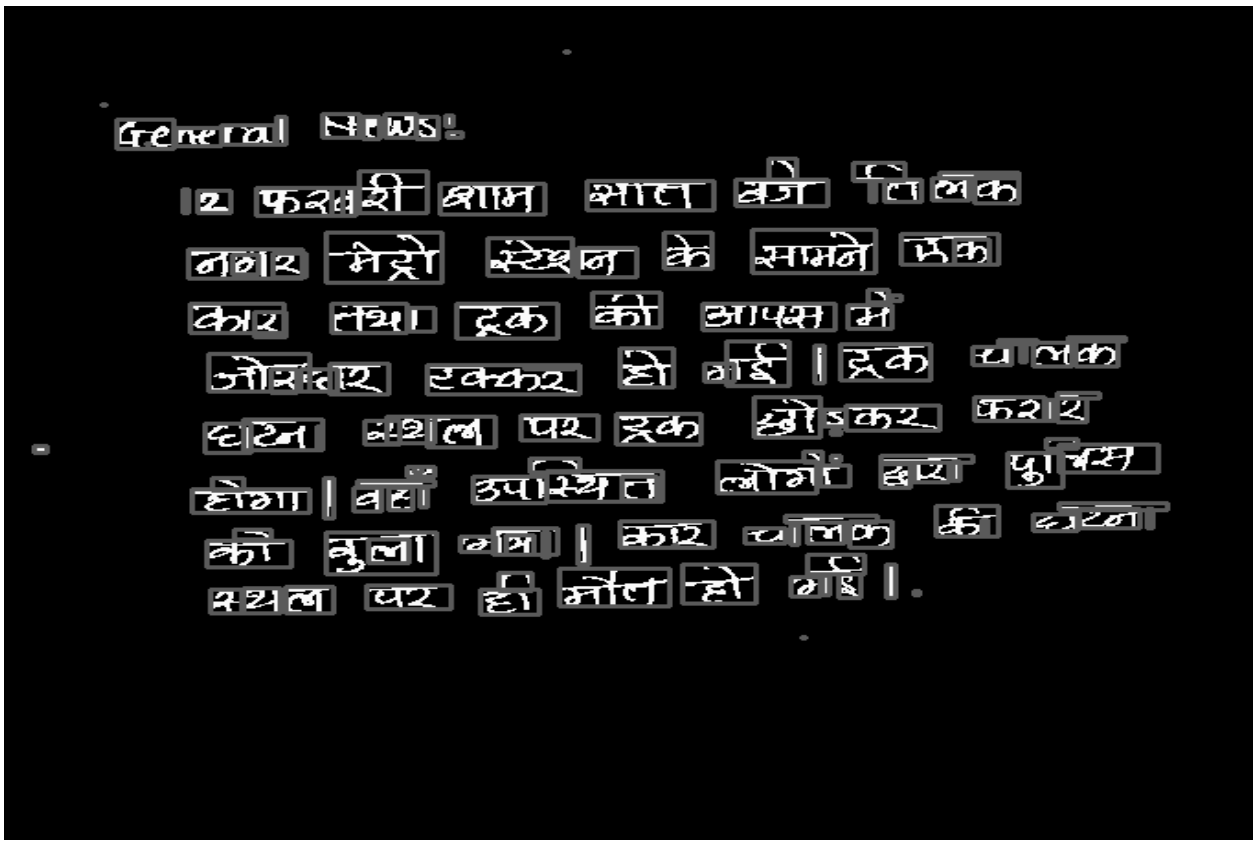


Preprocessing step: Images of contrast and gray scaling





Preprocessing step: - Images of binarisation and dilation.



Final image of segmentation.

Conclusions and Future Scope

Hereby we will be successfully implemented up to the segmentation step of optical character recognition model using machine learning.

This work presents a robust scheme of extracting text line from a scanned handwritten document image.

The ongoing research work on line and ligature based Kannada Text Line Extraction Based On Energy Minimization And Skew Correction systems motivated us to put forward more accurate segmentation algorithms for Hindi document images.

In future we will continue our work further on OCR model.

Bibliography

[1] A. Daud, W. Khan, and D. Che, “Urdu language processing: A survey,” Artif. Intell. Rev., vol. 47, no. 3, pp. 279–311, 2017.
[2] S. Naz, K. Hayat, M. I. Razzak, M. W. Anwar, S. A. Madani, and S. U. Khan, “The optical character recognition of Urdu-like cursive scripts,” Pattern Recognit., vol. 47, no. 3, pp. 1229–1248, 2014.
[3] G. S. Lehal and A. Rana, “Recognition of nastalique urdu ligatures,” in Proc. 4th Int. Workshop Multi-Lingual OCR, 2013, p. 7
[4] U. Pal and A. Sarkar, “Recognition of printed urdu script,” in Proc. ICDAR, 2003, pp. 1183–1187, 2003. [5] D. Satti and K. Saleem, “Complexities and implementation challenges in offline urdu nastaliq OCR,” in Proc. Conf. Lang. Technol., 2012, pp. 85–91.
[6] S. Hussain, “Complexity of Asian writing systems: A case study of Nafees Nasta’leeq for urdu,” in Proc. 12th AMIC Annu. Conf. e-Worlds, Governments, Bus. Civil Soc., Asian Media Inf. Center, Singapore, 2003. [7] S. T. Javed and S. Hussain, “Improving Nastalique specific pre-recognition process for Urdu OCR,” in Proc. IEEE 13th Int. Multitopic Conf. (INMIC), Dec. 2009, pp. 1–6.
[8] G. S. Lehal, “Ligature segmentation for urdu OCR,” in Proc. 12th Int. Conf. Document Anal. Recognit. (ICDAR), Aug. 2013, pp. 1130–1134.

Appendices

Detailed information, lengthy derivations, raw experimental observations etc. are to be presented in the separate appendices, which shall be numbered in Roman Capitals (e.g. Appendix I). Since reference can be drawn to published/unpublished literature in the ap- pendices these should precede the Literature Cited section.

Appendix-A: NS2 Download and Installation

1. Downloading and installing Anaconda 3
2. Double click the installer to launch.
3. Select an install for “Just Me” unless you’re installing for all users (which requires Windows Administrator privileges) and click Next.
4. Select a destination folder to install Anaconda and click the Next button.
5. Choose whether to add Anaconda to your PATH environment variable. We recommend not adding Anaconda to the PATH environment variable, since this can interfere with other software. Instead, use Anaconda software by opening Anaconda Navigator or the Anaconda Prompt from the Start Menu
6. Choose whether to register Anaconda as your default Python. Unless you plan on installing and running multiple versions of Anaconda, or multiple versions of Python, accept the default and leave this box checked.
7. If you wish to read more about Anaconda Cloud and how to get started with Anaconda, check the boxes “Learn more about Anaconda Cloud” and “Learn how to get started with Anaconda”. Click the Finish button.
8. After your install is complete, verify it by opening Anaconda Navigator, a program that is included with Anaconda: from your Windows Start menu, select the shortcut Anaconda Navigator. If Navigator opens, you have successfully installed Anaconda. If not, check that you completed each step above, then see our Help page

Installing opencv using anaconda prompt.

Open anaconda prompt and type the following in it

->> conda install –c menpo opencv

Installing Jupyter notebook from anaconda prompt

python3 –m pip install --upgrade pip
python3 –m pip install jupyter

Accessing jupyter notebook from anaconda prompt

->> jupyter notebook

Acknowledgement

We have great pleasure in presenting the report on **OCR Model for handwritten imsge.** We take this opportunity to express our sincere thanks towards our guide **Sofia Mujawar** Department of Computer, APSIT thane for providing the technical guidelines and suggesitions regarding line of work. We would like to express our gratitude towards his constant encouragement, support and guidance through the development of project.

We thank **Prof. Sachin Malave** Head of Department of Computer Engineering, APSIT for his encouragement during progress meeting and providing guidelines to write this report.

We also thank the entire staff of APSIT for their invaluable help rendered during the course of this work. We wish to express our deep gratitude towards all our colleagues of APSIT for their encouragement.

Student Name1: Anuja Velaskar
Student ID1: 1602042

Student Name2: Nidhi Munavalli
Student ID2: 16102049

Student Name3: Apurva Waingaonkar
Student ID3: 16102050

Publication

Paper entitled **"Kannada text line extraction based on energy minimization and skew correction"** is presented at **"International Conference/Journal Name"**
by **"Sunanda Dixit".**