Journal of Advanced Research in Image Processing and Applications
Volume 2, Issue 2 - 2019, Pg. No. 1-6
Peer Reviewed Journal

Article

Intelligent Document Parser with Zonal OCR and Text to Speech

Seema Kolkur[1], Rahul Joshi[2], Kadiri Saqlain[3], Joshi Sanket[4]

[1,2,3,4]Dept., of Computer Engineering, Thadomal Shahani Engineering College, Mumbai, India.

I N F O

Corresponding Author:
Seema Kolkur, Dept., of Computer Engineering, Thadomal Shahani Engineering College, Mumbai, India.
E-mail Id:
kolkur.seema@gmail.com
How to cite this article:
Kolkur S, Joshi R, Kadiri Saqlain K et al. Intelligent Document Parser with Zonal OCR and Text to Speech. *J Adv Res Image Proc Appl* 2019; 2(2): 1-6.

Date of Submission: 2020-01-06
Date of Acceptance: 2020-01-17

A B S T R A C T

Optical Character Recognition (OCR) is well researched and heavily used technology in the field of image processing and computer vision. With the advent of technology, we need to digitize different documents every day and hence OCR is used in many applications in recent years. It is a method of converting a scanned image into text. OCR looks at each line of the image and attempts to determine if the black and white dots represent a particular letter or number. However traditional OCR considers a document as a whole, but in some cases, we may be interested only in the specific part of the document. Zonal OCR is used for this purpose. Our system gives a solution to all document parsing needs. Along extraction of data it provides facilities of data sharing, spell checking and text to speech.

Keywords: OCR, Image Segmentation, Feature Extraction, Spell Checker

Introduction

OCR is a technology that enables a user to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data.[1] The technology deals with the problem of recognizing all different kinds of characters. Printed characters can be recognized and converted into machine readable text easily. Images captured by a digital camera differ from scanned documents or image-only PDFs. They often have defects such as distortion at the edges and dimmed light, making it difficult for most OCR applications to correctly recognize the text. But with the advent of technology, today most of document editing softwares offer out-of-the-box OCR capabilities.

However, this is not enough in some cases. For example, if you are not interested in the whole text of a document, but rather want to pull certain text elements which are located at specific positions. This is where "Zonal OCR" comes into play. It is also referred to as Template OCR. Zonal OCR basically allows extracting only the important data fields from a scanned document and store the extracted values in a structured database like excel. Zonal OCR can be used for automated data-entry of manually filled forms. This advanced and powerful OCR system will save a lot of time and effort when creating, processing and re-purposing various documents.

State of The Art

Figure 1, shows steps of general Optical Character Recognition (OCR) system.[2]

Steps of OCR System

Preprocessing

The aim of preprocessing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. Image pre-processing can significantly increase the reliability of an optical inspection.[3] Several filter operations which intensify or reduce certain image details enable an easier or faster evaluation. Users are able to optimize a camera image with just a few clicks. The techniques encompassed

in preprocessing are normalization, cropping, alignment, focus, selective focus, user-specific filter, static/ dynamic binarization, image plane separation and binning.[4,5]

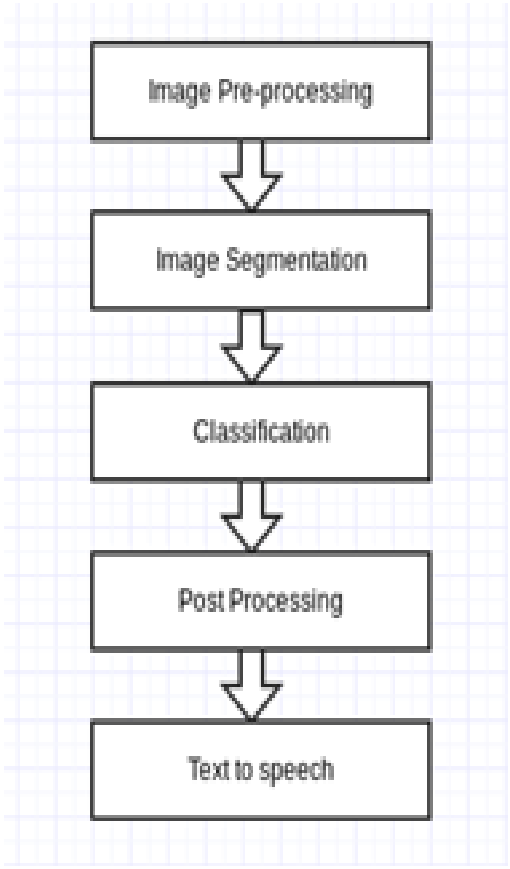


Figure 1.Steps of OCR

Image Segmentation

Segmentation is a process that determines the constituents of an image. It is necessary to locate the regions of the document where data have been printed and distinguish them from figures and graphics. For instance, when performing automatic mail-sorting, the address must be located and separated from other print on the envelope like stamps and company logos, prior to recognition. Applied to text, segmentation is the isolation of characters or words. The majority of optical character recognition algorithms segment the words into isolated characters which are recognized individually.[5]

Classification

Classification means assigning meaning to the segmented characters i.e. giving the recognized output result in the editable format. OCR systems require performing classification by a trained model like Artificial Neural Network (ANN). Tesseract uses a two-pass approach to character recognition. The second pass is known as "adaptive recognition" and uses the letter shapes recognized with high confidence on the first pass to recognize better the remaining letters on the second pass. This is advantageous for unusual fonts or low-quality scans where the font is distorted (e.g. blurred or faded). Tesseract has unicode (UTF-8) support, and can recognize more than 100 languages "out of the box". Tesseract can be trained to recognize other languages.[6,7]

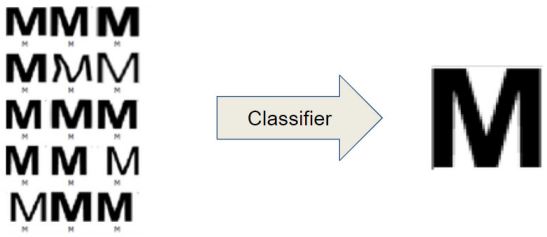


Figure 2.Classification of letter 'M'

Post Processing

The output of the classification process goes through an error detection and correction phase. This phase consists of the following three steps:

- Select an appropriate partition of the dictionary based on the characteristics of the input word, select the candidate words from the selected partition to match the input word with.
- Match the input word with the selected words.
- In case the input word is found in the dictionary, no more processing is done and the word is assumed to be correct. If the word is not found, there are two options available. We can generate aliases for the input word or restrict to an exact match.[16]

Existing Solutions

Some existing solutions are Tesseract, Google Cloud Vision, OCRopus, ABBYY, FineReader, etc.

- Tesseract was developed as proprietary software by Hewlett Packard Labs. It was open sourced in 2005. Since then Google and many open source developers contributed to its development. Tesseract V3.04, released in July 2015, supports over 100 languages. It can be used for more complicated OCR tasks like layout analysis by using a frontend such as OCRopus. Tesseract's output will have very poor quality if the input images are not preprocessed to suit it: Images(especially screenshots) must be scaled up such that the text x-height is at least 20 pixels. Any rotation or skew must be corrected or no text will be recognized, low-frequency changes in brightness must be high-pass filtered, or Tesseract's binarization stage will destroy much of the page, and dark borders must be manually removed, or they will be misinterpreted as characters.[7]
- OCRopus is a free document analysis and OCR system

released under the Apache License v2.0 with a very modular design using command-line interfaces. Once installed, it can be invoked by specifying the input images. It will output the recognized text to standard output directly or write it as hOCR (HTML-based) code into files, from which it then can be transformed to a searchable PDF. If more precise control is needed, options can be specified on the command line to perform specific operations.[8]

- ABBYY FineReader is an OCR application developed by the Russian company ABBYY. The program allows the conversion of image documents (photos, scans, PDF files) into editable electronic formats. In particular, Microsoft Word, Microsoft Excel, Microsoft PowerPoint, Rich Text Format, HTML, PDF/A, searchable PDF, CSV and txt(plain text) files.[9,10]
- Google Cloud Vision enables developers to build image recognition and classification features into the application, by incorporating image analytics capabilities in the form of easy to use REST APIs. Google Cloud Vision text recognition feature is able to detect a wide variety of languages and can detect multiple languages within a single image. Providing a language hint to the service is not required, but can be done if the service is having trouble detecting the language used in your image.

OCRdroid is a framework for digitizing images on smart phones.[11,12]

Methodology

Purpose of this work is to describe an application named Quick Form Lens as a full-fledged application for all document scanning and parsing needs. It allows setting up different types of Forms (templates) and based on the form which is set up, recognizes English Handwritten characters in the values field and stores them as information. The application takes scanned images of the form samples as input and store the values associated with different fields into electronic form. Also form data is speech out and/or can download in CSV format.

This application can be used for converting manually filled forms to digital data automatically. Some uses cases are patient healthcare records, identifying vehicle numbers from registration plates, processing cheques and signatures, invoice management and tracking, business card reader and retrieving Captcha. Figure 3 depicts flow diagram of the system.

Template Creation

A user can capture image of the (blank) form that he/she wants to digitize. In the application, a user can upload the image as new form template using pressing 'New'option. The form templates can be saved by giving appropriate name as shown in Figure 4. The user is then diverted to image cropping facility where all form field values should be enclosed in a box. Ucrop Cropping Library is used here with which a user can also work on other tasks like deskewing an image, rotate an image or changing aspect ratio of an image etc.[13] After getting final image, a user can selectively choose fields that need to be stored. User can type the name of the field and then click on the bottom left co-ordinate of its corresponding value field to get the box co-ordinate. In Figure 5a) co-ordinates of the field "project name" are displayed. These co-ordinated values (Figure 5b) are later used for key-value pairing in the sample detection process.

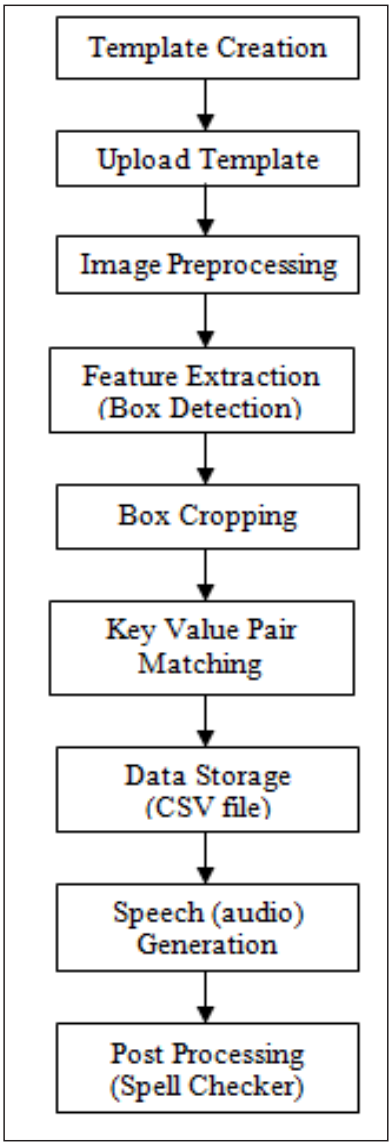


Figure 3.Flow diagram of the system

After entering all the field details, user can click on 'Create Form' button for setting up the form template in database. After the template is created, user can upload the respective samples from which information is to be extracted. Template creation is one-time process for every new template.

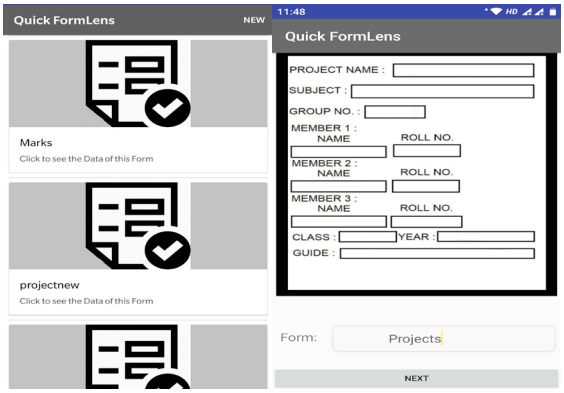


Figure 4.Template Creation

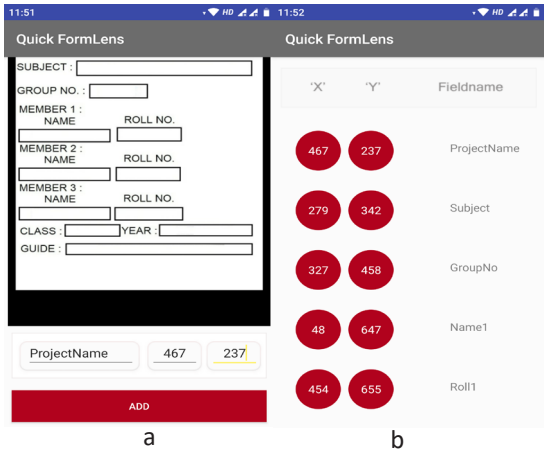


Figure 5.Finding Coordinate Values

Sample Detection

To detect new sample document (filled form), a user needs to navigate to its template and click on 'New' and upload the sample document image from the gallery. Sample document is cropped similar to actual template document.

Image Preprocessing

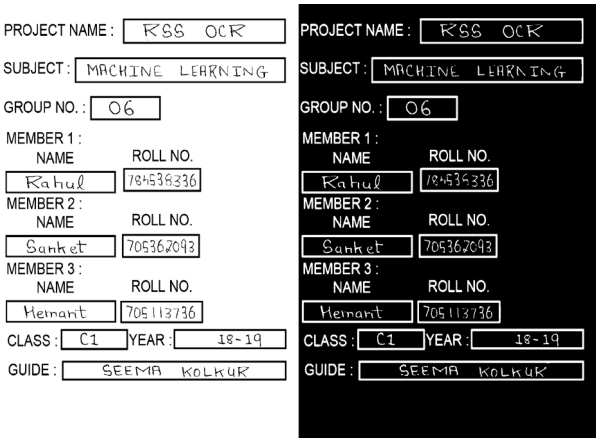
Original image is resized for better pixel quality. Otsu's thresholding method is used to obtain binary image for the input image (See Figure 6). The binary image is further inverted for further processing i.e. to obtain the cropped boxes.

Feature Extraction

Using image segmentation, horizontal and vertical lines are determined separately. These lines are then combined to detect boxes as shown in fig 7. All individual boxes are cropped as individual features.

Data Processing

Key-Value pairs of field names and detected boxes are matched to get values of individual fields as shown in Figure 8. The coordinates selected in the template are used to map various fields with their corresponding value in the current sample image. Sometimes, Key-Value pair needs to be re-matched to match the original.



(a) Original Binary Image (b) Inverted Image

Figure 6.Image Preprocessing

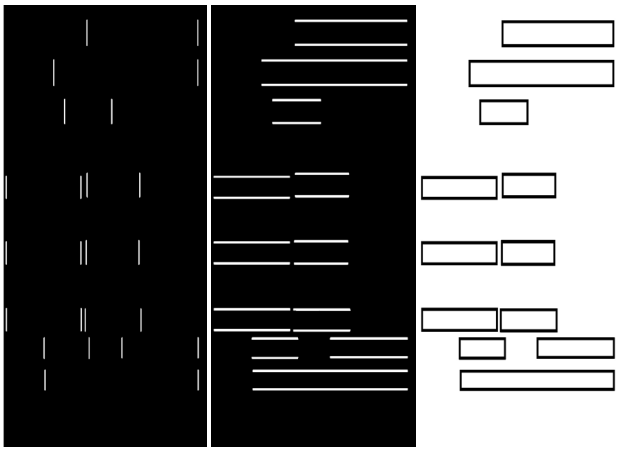


Figure 7.Feature Extraction

Post Processing

After capturing corresponding data for individual fields, there may be some error during scanning. Some words may be misspelled. This is handled by using spell checker which can be used by clicking "Suggest" button. A dataset of general English words and names of entities (students) in the form of a text file is used for correcting the misspelled words. These may be organization specific dictionaries. Using Norvig Spell Checker[14] and words dataset, suggestions can be provided for each incorrect field. For Eg for word 'Kahul' appropriate word 'Rahul' is suggested (Figure 9). Multiple suggestions may be given. The dataset consists of words which contains frequently used values in different fields from multiple templates provided to the system. After using this dataset with our spell checking algorithm, accuracy of spell check process was increased (Table 1).

Table 1. Accuracy of Spell Checker

Accuracy	Without Spell Checker	With Spell Checker
Printed	84%	86%
Hand-Written	65%	72%

Figure 8.Data after Key-Value Pair matching

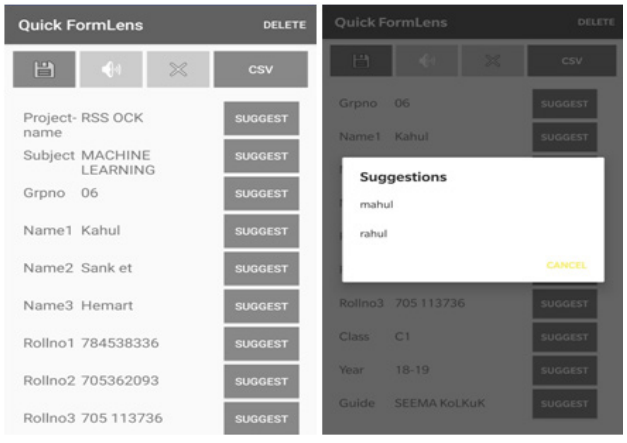


Figure 9.Suggestions after spell checker

Data Storage and Sharing

All the extracted data can be made persistent by storing it in a CSV file just a single click. It can also be shared using Whatsapp, Gmail etc as shown in figure 10.

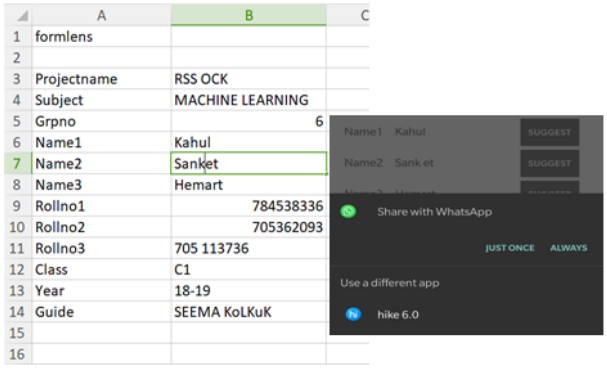


Figure 10.Data Storage and Sharing

Speech and Media Player

Sometimes the extracted data may be required to speak out for example to assist blind people in their day to day activities. The file can be stored for future reference in the form of an audio (.wav) file using the record button (Figure 11). Play pause buttons are provided for user-friendly control of audio output.[16]

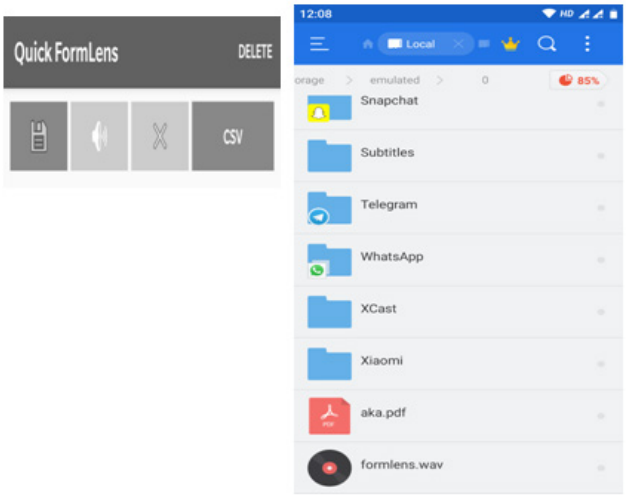


Figure 11.Recording Audio file

Conclusion

Simple digitization of documents is not sufficient for some applications. OCR helps to convert different types of brochures into editable and searchable data. Zonal OCR is used when we are interesting into specific parts of a document. An effective document parser system is required to fully utilize all benefits after OCR. Because of inherent limitations of OCR system, there may be errors in data extraction. We can overcome this by using spell checker during post-processing. We also need to store and share data. To aid the blind people this data can be rehabilitated to audio file. All these facilities are provided in our system. It acts as an end-to-end software for all data parsing activities. It provides satisfactory results. The system can be improved further by ssupporting multiple

languages for text to speech like English text to Hindi etc. Also multiple languages for OCR can be supported like extracting Hindi Text. We used enhanced image processing for better accuracy by use of different algorithms.

References

1. ARCHANA A. SHINDE, D. 2012, "Text Pre-processing and Text Segmentation for OCR", International Journal of Computer Science Engineering and Technology, pp. 810- 812.
2. Mithe, Ravina, Supriya Indalkar, and Nilam Divekar. "Optical character recognition." International journal of recent technology and engineering (IJRTE) 2.1 (2013): 72-75.
3. Bieniecki, Wojciech et al. "Image Preprocessing for Improving OCR Accuracy." 2007 International Conference on Perspective Technologies and Methods in MEMS Design (2007): 75-80.
4. Devi, H.K.A., 2006. "Thresholding: A Pixel-Level Image Processing Methodology Preprocessing Technique for an OCR System for the Brahmi Script", Ancient Asia, 1, pp.161–165.
5. W. Bieniecki, S. Grabowski, and W. Rozenberg,"Image preprocessing for improving ocr accuracy", Perspective Technologies and Methods in MEMS Design, MEMSTECH 2007.
6. Smith, Ray. "An overview of the Tesseract OCR engine." Ninth International Conference on Document Analysis and Recognition (ICDAR 2007). Vol. 2. IEEE, 2007.
7. Patel, Chirag, Atul Patel, and Dharmendra Patel. "Optical character recognition by open source OCR tool tesseract: A case study." International Journal of Computer Applications 55.10 (2012): 50-56.
8. Breuel, Thomas M. The OCRopus open source OCR system. *Electronic Imaging* (2008).
9. OCR, PDF, Text Scanning, Software and Solutions – ABBYY"https://www.abbyy.com/en-apac/"
10. Abbyy Mobile OCR Engine. http://www.abbyy.com/mobileocr/.
11 Zhang, Mi & Joshi, Anand & Kadmawala, Ritesh & Dantu, Karthik & Poduri, Sameera & Sukhatme, Gaurav. (2019). OCRdroid: A Framework to Digitize Text on Smart Phones.
12. Sonia Bhaskar, Nicholas Lavassar, Scott Green. Implementing Optical Character Recognition on the Android Operating System for Business Cards", EE 368 Digital Image Processing.
13. https://github.com/Yalantis/uCrop last retrived on 16/12/2019
14. Khushbu, Isha Vats, Image Segmentation Algorithm: A Review,IJIRCCE Vol. 5, Issue 6, June 2017.
15. How to Write a Spelling Correctorhttp://norvig.com/spell-correct.html
16. D.Sasirekha, E.Chandra, "Text To Speech: A Simple Tutorial", March 2012.