Simon Fong · Nilanjan Dey · Amit Joshi
Editors

ICT Analysis and Applications

Springer

Editors

Simon Fong
Department of Computer and Information Science
University of Macau
Macau, Macao

Nilanjan Dey
JIS University
Kolkata, India

Amit Joshi
Global Knowledge Research Foundation
Ahmedabad, India

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-981-16-5654-5 ISBN 978-981-16-5655-2 (eBook)
https://doi.org/10.1007/978-981-16-5655-2

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

A Detailed Review on Text Extraction Using Optical Character Recognition

Chhanam Thorat, Aishwarya Bhat, Padmaja Sawant, Isha Bartakke, and Swati Shirsath

Abstract There exist businesses and applications that involve huge amount of data generated be it in any form to be processed & stored on daily basis. It is an implicit requirement to be able to carry out quick search through this enormous data in order to deal with the high amount of document and data generated. Documents are being digitized in all possible fields as collecting the required data from these documents manually is very time consuming as well as a tedious task. We have been able to save a huge amount of efforts in creating, processing, and saving scanned documents using OCR. It proves to be very efficient due to its use in variety of applications in Healthcare, Education, Banking, Insurance industries, etc. There exists sufficient researches and papers that describe the methods for converting the data residing in the documents into machine readable form. This paper describes a detailed overview of general extraction methods from different types of documents with different forms of data and in addition to this, we have also illustrated on various OCR platforms. The current study is expected to advance OCR research, providing better understanding and assist researchers to determine which method is ideal for OCR.

Keywords Optical character recognition · Data extraction · Data pre- processing · Segmentation · Classification · Post processing · Feature extraction · Neural networks

C. Thorat (✉) · A. Bhat · P. Sawant · I. Bartakke · S. Shirsath
MKSSS's Cummins College of Engineering for Women, Pune, India
e-mail: chhanam.thorat@cumminscollege.in

A. Bhat
e-mail: aishwarya.bhat@cumminscollege.in

P. Sawant
e-mail: padmaja.sawant@cumminscollege.in

I. Bartakke
e-mail: isha.bartakke@cumminscollege.in

S. Shirsath
e-mail: swati.shirsath@cumminscollege.in

S. Fong et al. (eds.), *ICT Analysis and Applications*, Lecture Notes
in Networks and Systems 314, https://doi.org/10.1007/978-981-16-5655-2_69

1 Introduction

Businesses are expanding and updating as technology is advancing day by day making life and work easier to deal with. As this process tends to be a tedious task & data is in huge amount, it leads to excessive manual work. In order to automate this process of data collection and entry, Optical Character Recognition was introduced. It is used to recognize the text that is required to be extracted from scanned documents such as pdf, jpeg, handwritten etc. and convert it into machine readable text and store it in some target format. OCR allows this data to be quickly searched, edited, stored & displayed efficiently and that too digitally saving time as well as efforts. The capabilities of OCR provide endless possibilities throughout the entire industrial spectrum covering banking, healthcare, Education, Legal etc. thereby saving time and physical work required. Thus to gain insight from the data present in different kinds of documents and automatically extract the necessary data, OCR plays a key role and how it is done is thoroughly reviewed in the proposed paper.

2 OCR Process Flow

Building an OCR engine from the ground up is a step-by-step procedure that OCR professionals work on. The training process for an algorithm for efficient problem-solving with the use of optical character recognition typically involves six steps as shown in diagram (Fig. 1).

2.1 Image Acquisition

During the image acquisition phase, the recognition system collects a scanned image as an input image. The image must be in one of the supported formats, such as JPEG, PNG, or BMT. This image can be captured with a scanner, digital camera, or any other compatible digital input device.

2.2 Pre-processing

Pre-processing is a set of processes that are applied to the scanned input image. It basically, improves the image, making it segmentation-ready. In pre-processing, the image is subjected to a variety of activities. Binarization is one of them, which uses the global thresholding approach to convert a grayscale image to a binary image. The sobel technique is used to dilate the edges of the binarized image, and the image is dilated and filled with holes in the final two phases to generate the segmentation-ready image.

2.3 Segmentation

During the segmentation stage, an image of a series of characters is divided into sub-images of individual characters. The input image is separated into discrete characters after it has been pre-processed. in the OCR system by employing a labelling procedure to assign a number to each character. The amount of characters in the image is indicated by this labelling. Each character is scaled to a standard size of 30 × 20 pixels.

2.4 Classification and Recognition

The decision-making stage of the OCR recognition system is the classification stage. This research employs a feed forward back propagation neural network to classify and recognise handwritten characters. The classifier is fed 600 pixels from the resized character in the segmentation stage. The neural classifier has two hidden layers in addition to an input and output layer. The log sigmoid activation function is used in the hidden layers, and the output layer is a competitive layer because one of the characters must be identified at all times. The output layer has a total of 26 neurons because the OCR system is designed to recognise English alphabets.

2.5 Post-processing

The post-processing stage is the final stage of the OCR recognition system. It computes the equivalent ASCII value using the test samples' recognition index and showcases the relevant identified characters in structured textual format.

3 OCR Methods

Many researches have been conducted in an attempt to recognize characters, forming varied combinations of techniques to achieve higher accuracy. The accuracy is dependent on multiple factors including quality of source, page layout, content type, OCR technique and various other factors. In some cases, pre-processing that involves eliminating shadows, background noises, adjusting skewness, etc. is a necessary stage prior to OCR. Among a wide variety of methods [2] for performing OCR, the prominent ones are listed below.

3.1 Matrix Matching

A matrix is created for each character to find a particular pattern. This pattern is then compared with all the indexes of known characters. It is the perfect fit for one column layout document.

3.2 Fuzzy Logic

This strategy generates an output with multiple values in contrast to the binary values with only two values 0 or 1. Fuzzy logic signifies a more humanistic approach to logical thinking rather than binary-based computers. It makes an approximation between two characters to find similarities. An input with no noise is appropriate for this technique, for e.g. PDF document.

3.3 Extraction of features

Height, width, density, lines, loops, stems and some additional characteristics are the essential features involved in a character formation. Every character is constructed by mentioning presence or absence of these essential features. Mostly magazines, laser print, and high-quality images have such characters and hence this technique is useful there.

3.4 Structural Analysis

In this method, the image shape, sub-vertical and sub-horizontal histograms help to detect the characters. Structural Analysis has an exceptional potential to repair damaged characters. This makes it best suited for OCR on low-quality text and newsprints.

3.5 Neural Networks

Neural networks are based on human neural system. The pixels in an image are compared to an index of familiar character pixel patterns. Additionally, it has a great character abstraction capability ideal for faxed documents or distorted text.

Convolutional Neural network (CNN) and Long-short term memory (LSTM) [3] are some of the algorithms that help in recognition. Neural networks can be further utilized for analytical tasks.

4 Data Extraction from Specified Documents

As documents generated belong to different forms according to the need and usage, based on the research work done, this chronicle illustrates how data is extracted from the respective formats or type of document.

4.1 Exam Answer Sheets

The research paper [4] is a narrative of a unique approach for developing an automated, fast, flexible, and reliable system that can quickly recognize a student's enrollment number and related marks from an exam answer sheet and store it in a host computer. It makes use of hardware that takes each answer script's front page and processes it one at a time from the bundle. The entire procedure starts with preprocessing of the image which is captured by the Video Graphics Array (VGA) camera by using an adaptive noise removal algorithm, followed by the adaptive threshold for color detection algorithm which captures the roll number and marks from the answer sheet. The required data i.e. roll number and marks of the students are then passed on to the OCR system which converts it into a machine-readable format and generates an excel sheet of the same. As all educational organizations generate a large amount of student id and mark data, the proposed system helps reduce the manual work the staff is put through for the collection of marks from these exam sheets and saves up an adequate amount of time and effort required to do so. Integrating OCR will help enhance any Data entry system.

4.2 Print Bills and Invoices

The study [5] presents a methodology for extracting any data from printed bills and invoices, which may then be employed in machine learning or statistical analysis. It primarily focuses on extracting bill amount, schedule, date, and similar data in order to obtain sufficient information about the user's purchases and transactions, as well as likes and dislikes. Initially, the invoice or bill from the image is first detected using the edge detection algorithm and then the unwanted noise from the image is filtered out before the text is extracted using Text Segmentation. Data in tabular format may be included in the bills or invoices. Automatically extracting data from tables is a challenging task in itself, mainly because tables being in diverse formats. These tables have no standard table design format, different table layouts, rows, columns, cell structure schemes (single, merged, etc.), content alignments, spacing, font sizes, etc. A large number of table detection algorithms are present that detect the table boundary and determine the table structure in a document. Apart from OCR other table detection methods are X–Y cut, tag classification, etc. Detection becomes difficult when there is no table border or outer ruling present. The results from the table recognition can be used to execute queries in the future.

4.3 Newspaper & Comics

This proposed system [6] detects and extracts the text in images automatically which has been collected from varied sources like newspapers, videos, ads, photographs, and checks. Text is detected using a multiscale texture segmentation approach, followed by spatial cohesion restrictions, clean-up, and ultimately extraction using a histogram-based binarization approach. A system for automatically evaluating performance is also proposed. The technique works well for standard documents as well as those with non-structured layouts or those with text written in shaded or textured conditions.

4.4 Multilingual Documents

The majority of accessible OCR engines are limited to a single language or a small number of languages. When different languages are present in a single text, additional challenges arise. Consider a document that has both English and Hindi text. As a result, the following processing stages are proposed for a typical multilingual OCR system:

- Binarization, layout analysis, and page segmentation are the first steps in pre-processing, followed by line finding and, if desired, word and character segmentation.
- Last but not least, feature extraction and classification or recognition.

This paper [7] also shows an automatic script identification methodology for documents, which is divided into two categories: segmentation-based and free techniques, depending on whether text lines are subdivided into finer units before recognition. To discover the best match, segmentation-based approaches break a text line into a number of smaller units made up of characters or sub-characters and compare a combination of these units to likely lexical words. In contrast, segmentation-free OCR approaches, such as the left-to-right hidden Markov model, implicitly integrate character segmentation in producing a globally optimal character/word series as the recognition result (HMM). Different recognition modelling strategies can be utilized depending on whether a segmentation technique is segmentation-based or free.

5 OCR Platforms

With the introduction and expansion of computerised systems, many data scientists have attempted to solve the problem of OCR difficulties by developing a practical and efficient OCR engine that can produce accurate results. As we all know, there are a variety of OCR techniques and platforms accessible today, but we're just going to look at a few.

5.1 Google Docs OCR

It is easily available OCR facility proposed by Google within the Google Drive service where one can upload [9] an image file or a scanned document to Google Docs then it will automatically perform OCR on it generating an output file in any of the following formats such as PDF, ODT, TXT, DOC, RFT and HTML. Apart from that it also supports 30 languages.

5.2 Tesseract OCR

It's an open source OCR engine for a variety of operating systems. It includes an OCR engine called libtesseract as well as a command-line tool called tesseract. It also focuses on recognition with the addition of a new neural set (LSTM) based OCR service, Unicode (UTF-8) support, and the capacity to recognise over 100 languages. Tesseract supports a variety of output formats, including plain text, PDF, OCR (HTML), PDF with invisible text only, and TSV.

5.3 ABBYY FineReader

An international company named "ABBYY" designed and developed this OCR system which provides efficient OCR services. Prior to performing OCR on document, the program analyses the structure of the entire document and identifies the region that comprises of the text or any images, tables and barcodes [8]. The results achieved through recognition are then displayed in the text window where unknown characters are emphasized in the window which helps the user to track down likely possible errors, figure them out and rapidly correct them within the software.

5.4 Transym

Another OCR solution that aids R&D organisations in obtaining precise and accurate information from digitised documents is Transym. The developed system is offered as an SDK with a high-level API, as well as a software package with a simple GUI that can be quickly installed and utilised.

5.5 I2OCR

It is a freely available online OCR tool which first extracts text from images and then converts it into text format which can be later edited, formatted, indexed, searched, or translated. The supports input images which are of following file types i.e. TIF, JPEG, BMP, PNG, GIF, PGM, PBM and PPM. It also supports more than 60 languages along with multi column analysis.

6 Future Work

OCR can read Screenshots supporting the information transfer between incompatible technologies, for example - .jpg and.docx. It is fascinating when combined with artificial intelligence to auto-extract information from scanned receipts to generate an expense report, or to translate a foreign language using a phone’s camera. The future encompasses systems that first recognise the scanned text and then derives meaning from it. It can also facilitate development in the field of robotics. Robots comprehending text opens up a wide range of applications. Furthermore, OCR can be integrated in anti-viruses to catch viruses by detecting codes hidden in images.

7 Conclusion

There are many advancements in OCR that are still under development but this review paper discusses how data is automatically extracted from different types of documents be it an image or pdf, a handwritten document, multilingual documents,

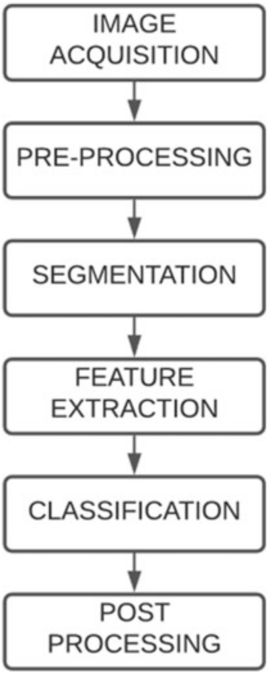

Fig. 1 Optical Character Recognition Process Flow [1]

invoices etc. using OCR. A generic methodology that goes in to make OCR work is also reviewed. The paper presents a brief survey of different OCR platforms that are available and have been used. This paper will act as a good guide for researchers preceding their study in the field of OCR.

References

1. J. Pradeep, E. Srinivasan, S. Himavathi, Neural network based handwritten character recognition system without feature extraction, in *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)*, pp. 40–44 (2011). https://doi.org/10.1109/ICCCET.2011.5762513
2. S. Singh, Optical character recognition techniques: A survey. (IJARCET) **2**(6) (June 2013)
3. R. Mittal, A. Garg, Text extraction using OCR: A systematic review, in *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 357–362 (2020). https://doi.org/10.1109/ICIRCA48905.2020.9183326
4. D. Sharma, A. Sharan, H. Sharma, A. Agarwal, Data extraction from exam answer sheets using OCR with adaptive calibration of environmental threshold parameters, pp. 498–502 (2013). https://doi.org/10.1109/ICSPCom.2013.6719843
5. H. Sidhwa, S. Kulshrestha, S. Malhotra, S. Virmani, Text extraction from bills and invoices, in *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, pp. 564–568 (2018). https://doi.org/10.1109/ICACCCN.2018.8748309

6. V. Wu, R. Manmatha, E. M. Riseman, Textfinder: an automatic system to detect and recognize text in images, IEEE Trans. Pattern Anal. Mach. Intell. **21**(11), 1224–1229 (Nov 1999). https://doi.org/10.1109/34
7. X. Peng, H. Cao, S. Setlur, V. Govindaraju, P. Natarajan, Multilingual OCR research and applications: an overview. *MOCR '13* (2013)
8. A. Mehdi, OCR as a service: A experimental evaluation of google docs OCR, Tesseract, ABBYY FineReader, and Transym. ISCV. Springer International Publishing AG, pp. 66 (2016) https://doi.org/10.1007/978-3-319-50835-1
9. S. Vijayarani, A. Sakila, Performance comparison of OCR tools. Int. J. UbiComp. **6**, 19 (2015). https://doi.org/10.5121/iju.2015.6303