IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

Building Text Extraction using OCR

Mayur Zagade[1], Shivani More[2], Manish Pasalkar[3], Anand Narute[4], Prof. Anuradha Thorat[5]
Students, Department of Information Technology[1,2,3,4]
Assistant Professor, Department of Information Technology[5]
Zeal College of Engineering and Research, Pune, Maharashtra, India

Abstract: *Image recognition and optical character recognition technologies have become an integral part of our everyday life due in part to the ever-increasing power of computing and the ubiquity of scanning devices. Printed documents can be quickly converted into digital text files through optical character recognition and then be edited by the user. Consequently, minimal time is required to digitize documents; this is particularly helpful when archiving volumes of printed materials. This study demonstrates how image-processing technologies can be used in combination with optical character recognition to improve recognition accuracy and to improve the efficiency of extracting text from images. Two software systems are developed and tested during this study: a character recognition system applied to cosmetic-related advertising images and a text detection and recognition system for natural scenes. The results of the experiment demonstrate that the proposed systems can accurately recognize text in images.*

Keywords: Cosmetic, Ubiquity, Optical

I. INTRODUCTION

Attempts have long been made to design computer programs that can read printed documents with the objective of improving archiving efficiency by converting documents into electronic files in an automated manner. Systems capable of recognizing text in images and converting it into characters for editing on a computer are known as optical character recognition (OCR) systems . OCR was first proposed by the German scientist Tausheck in 1929. Since the 1960s, scientists worldwide have sought to improve OCR using computers. Early OCR research was focused on identifying the numerals 0–9. The earliest research on recognition of printed Chinese characters was conducted by Casey and Nagy, who published their first paper on Chinese character recognition in 1966; this paper details the successful identification of 1,000 printed Chinese characters by using a template matching technique. In this paper, we discuss our development of two OCR based systems: a character recognition system for commercial advertising images and a text detection and recognition system for natural scenes. After the basic character recognition system is completed, we will integrate an improper words detection system to preemptively reduce the number of legal disputes that can arise from using inappropriate words in advertisements. The main purpose of the text recognition system to be applied to natural scenes is to assist managers in archiving documents.

II. RELATED WORK

Assumptions and Dependencies

Assumption:

In this system, we take an image dataset as input and, using the OCR algorithm , And Recognize text from Images.

Dependencies:

[1] Used Python Language:Python is commonly used for developing websites and software, task automation, data analysis, and data visualization.

[2] Since it's relatively easy to learn, Python has been adopted by many non-programmers such as accountants and scientists, for a variety of everyday tasks, like organizing finances.Python is a general-purposeprogramming language, so it can be used for many things.

[3] Python is used for web development, AI, machine learning, operating systems, mobile application development, and video games. ... Python is a relatively easy programming language to learn and follows an organized structure The python language is one of the most accessible programming languages available because it has simplified syntax and not complicated, which gives more emphasis on natural language.

[4] Due to its ease of learning and usage, python codes can be easily written and executed much faster than other

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

programming languages. Machine learning: Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so.

[5] Machine learning algorithms use historical data as input to predict new output values. Machine learning is a method of data analysis that automates analytical model building.

[6] It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.

III. SYSTEM ARCHITECTURE

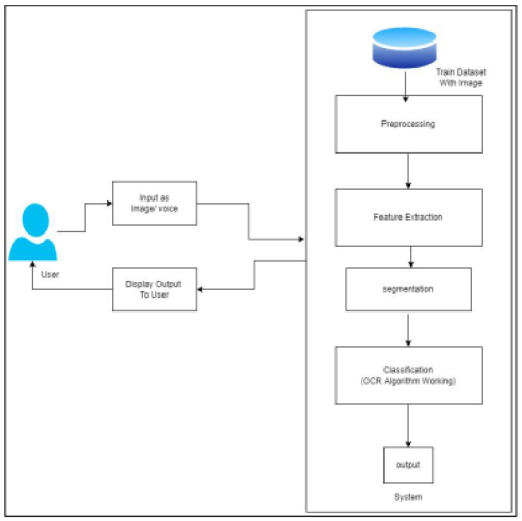

Fig: - System Architecture

[1] Requirement Analysis - Requirement Analysis is the most important and necessary stage in SDLC. The senior members of the team perform it with inputs from all the stakeholders and domain experts or SMEs in the industry. Planning for the quality assurance requirements and identifications of the risks associated with the projects is also done at this stage. Business analyst and Project organizer set up a meeting with the client to gather all the data like what the customer wants to build, who will be the end user, what is the objective of the product. Before creating a product, a core understanding or knowledge of the product is very necessary.

[2] System Design - The next phase is about to bring down all the knowledge of requirements, analysis, and design of the software project. This phase is the product of the last two, like inputs from the customer and requirement gathering. [3]. Implementation - In this phase of SDLC, the actual development begins, and the programming is built. The implementation of design begins concerning writing code. Developers have to follow the coding guidelines described by their management and programming tools like compilers, interpreters, debuggers, etc. are used to develop and implement the code.

[4] Testing - After the code is generated, it is tested against the requirements to make sure that the products are solving the needs addressed and gathered during the requirements stage. During this stage, unit testing, integration testing, system testing, acceptance testing are done.

[5] Deployment - Once the software is certified, and no bugs or errors are staSted, then it is deployed. Then based on the assessment, the software may be released as it is or with suggested enhancement in the object segment. After the software is deployed, then its maintenance begins.

[6] Maintenance - Once when the client starts using the developed systems, then the real issues come up and requirements to be solved from time to time. This procedure where the care is taken for the developed product is known as maintenance

IV. CONCLUSION

Need several kinds of images as sources of information for elucidation and analysis. When an image is transformed from one form to another such as digitizing, scanning, and communicating, storing, etc. degradation occurs. Therefore, the output image has to undertake a process called image enhancement, which contains of a group of methods that seek to develop the visual presence of an image. Image enhancement is fundamentally enlightening the interpretability or awareness of information in images for human listeners and providing better input for other automatic image processing systems. New features can be added to improve the accuracy of recognition. These algorithms can be tried on large database of handwritten text. There is a need to develop the standard database for recognition of text. The proposed work can be extended to work on degraded text or broken characters. Recognition of digits in the text, half characters and compound characters can be done to improve the word recognition rate. This extracted text can be further converted to audio so make physically challenged i.e. blind people easily understand which text has been converted from the image.

REFERENCES

[1]. X. Liu, D. Liang, S. Yan, D. Chen, Y. Qiao, and J. Yan, “FOTS: Fast oriented text spotting with a unified network,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 5676–5685.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

[2]. S. Unar, A. H. Jalbani, M. M. Jawaid, M. Shaikh, and A. A. Chandio, "Artificial urdu text detection and localization from individual video frames," Mehran Univ. Res. J. Eng. Technol., vol. 37, no. 2, pp. 429–438, 2018.

[3]. A. Mirza, M. Fayyaz, Z. Seher, and I. Siddiqi, "Urdu caption text detection using textural features," in Proc. 2nd Medit. Conf. Pattern Recognit. Artif. Intell., 2018, pp. 70–75.

[4]. C. Yao. MSRA Text Detection 500 Database (MSRA-TD500). Accessed: Aug. 2018 [Online].

[5]. A. A. Chandio and M. Pickering, "Convolutional feature fusion for multilanguage text detection in natural scene images," in Proc. 2nd Int. Conf. Comput., Math. Eng. Technol. (iCoMET), Jan. 2019, pp. 1–6.

[6]. A. A. Chandio and M. Pickering, "Convolutional feature fusion for multilanguage text detection in natural scene images," in Proc. 2nd Int. Conf. Comput., Math. Eng. Technol. (iCoMET), Jan. 2019, pp.1–6.