

Text Extraction and Recommendation System

Akshata S. Babar¹ Pranjal D. Mane² Rahul B. Jadhav³ Priyanka G. Kamble⁴

^{1,2,3,4}B.Tech. Student

^{1,2,3,4}AITRC, Vita, India

Abstract— The text extraction framework includes clever calculations to distinguish and separate the printed content present in different sorts of pictures. With the appearance of the computerized period and the accessibility of bunch of media substance, it has become critical to peruse and decipher the texts related with those substance. The programmed extraction of texts would not just serve to deduce the semantics of those interactive media archives yet additionally help in productive ordering and ensuing recovery of the equivalent. Nonetheless, the text contrasts in size, style, arrangement and so on and low goal of the foundation of complex pictures cause the issue of text to distinguish cation a complex one. Subsequently, the extraction of text information in pictures has turned into a difficult field of research in the area of Image Processing. The fundamental impediment of the current procedures, for example, surface based or associated segment based is that they are incapable to furnish exact outcomes with incredible accuracy for the utilizations of text extraction. The proposed Text Extraction System would cleverly peruse the text.

Keywords: Text Extraction

I. INTRODUCTION

With the data age progressing at a lightning pace, the media content in the advanced world is developing at an outstanding rate. Out of which, the picture and video information are generally common. Each second observers a great many pictures and recordings being caught by gadgets going from cell phones and individual cameras to reconnaissance cameras and satellite frameworks. Among every one of these pictures, a greater part of them contains text information installed in them. This text for the most part used to allude to the semantics of the compartment picture. For example, it can give an outline of the depiction for the picture like where the picture is taken, at what time and what subjects are available in the picture, etc thus forward. A particularly broad volume of picture information would be delivered pointless in many faculties if their semantics are not mined as expected. Henceforth, perceiving those inserted texts are of central significance since it will consequently assist with unraveling semantics of the picture.

A. Background

Character acknowledgment is anything but another issue however its underlying foundations can be followed back to frameworks previously the creations of PCs. The soonest OCR frameworks were not PCs but rather mechanical gadgets that had the option to perceive characters, yet exceptionally sluggish speed and low exactness. In 1951, M. Sheppard invented a reading and robot GISMO that can be considered as the most punctual work on present day OCR. GISMO can peruse melodic documentations just as words on a printed page individually. In any case, it can just perceive 23 characters. The machine additionally has the capacity to could duplicate a typewritten page. J. Rainbow, in 1954,

contrived a machine that can peruse capitalized typewritten English characters, one every moment. The early OCR frameworks were reprimanded because of mistakes and slow acknowledgment speed. Henceforth, very little exploration endeavors were put on the subject during 60's and 70's. The solitary advancements were done on government offices and huge companies like banks, papers and carriers and so forth As a result of the intricacies related with acknowledgment, it was felt that three ought to be normalized OCR text styles for facilitating the undertaking of acknowledgment for OCR. During the previous thirty years, substantial research has been done on OCR. This has lead to the rise of record picture examination (DIA), multi-lingual, transcribed and omni-textual style OCRs. In spite of these broad exploration endeavors, the machine's capacity to dependably read text is still far underneath the human. Consequently, momentum OCR research is being done on improving precision, speed of OCR for assorted style archives printed and written in unconstrained conditions. There has not accessibility of any open source or any business programming accessible for the complex language like Urdu or Sindhi.

B. Application in General Forms in Different Areas

Optical Character Recognition (OCR) is a piece of programming that converts printed text and pictures into digitized structure with the end goal that it tends to be controlled by machine. Dissimilar to human mind which has the ability to handily perceive the text/characters from a picture, machines are not keen enough to see the data accessible in picture. Consequently, countless exploration endeavors have been set forward that endeavors to change a report picture to design reasonable for machine.

C. Challenges in Existing System

For great quality and high exactness character acknowledgment, OCR strategies anticipate top caliber or then again high goal pictures for certain essential underlying properties, for example, high separating message and foundation. The manner in which pictures are produced is a significant and deciding component in the precision and accomplishment of OCR, since this frequently influences the nature of pictures significantly. Generally OCR with pictures delivered by scanners gives high accuracy and great performance. Conversely, pictures created by cameras for the most part are not as great of a contribution as filtered pictures to be utilized for OCR due to the natural or camera related components.

II. LITERATURE SURVEY

A. Existing Work

Tesseract is an open source optical character recognition system. It was developed at HP in between 1984 to 1994. It was modified and improved in 1995 with Great accuracy. In late 2005, HP release Tesseract for open source. It is now available at. It is highly portable. It is more focused towards

providing less rejection than accuracy. Currently only command base version is available. HP never used it. Now it is developed and maintained by Google. It provides support for various languages.

B. Theory of Work

Optical character Recognition (OCR) is a convert into scanning or printe the text from images handwritten text into editable text for further processing. This technology allows machines to recognize the text automatically. It is like a combination of eye and mind of the human body. An eye can see the text from the image but actually the brain processes and extracted text read by eye. CBDR is unquestionably an approach designed to alleviate the data management plus the user has the ability to access the data easily. Subsequently in content-based data recommendation, the extraction of an data in accordance with similarities in their features such as colors, textures, shapes and such like. These features are lower-level contents of an image.

C. Proposed Work Converted to Concept

In this system we extract the text from image and then provide the recommendations to users related the extracted text. Basically this system is divided in two parts, first is Text extraction system and second part is Text recommendation system. In text extraction system we are taking image as input, which consist of some text then we extract text from image. In this part we are using OCR (Optical Character Recognition). Texts usually have different appearance like font, size, style, orientation, alignment, texture, color, contrast, and background. All the changes will produce the problem for automatic text extraction become very complicated and difficult. In text recommendation system we are taking text from image and a send it to our system. In this part we are using CBDR (Content Based Data Recommendation). We are taking one image at a time .In this part we can provide recommendations to users related tot that test.

III. PURPOSE

Now days lots of data available in the form of image. People prefers the capturing data instead of writing somewhere and also there is no recommendation system available which provides recommendation based on the image. So proposed system will extract the text from image and provides appropriate recommendations to users. It will saves the time and increases performance of current recommendation system.

IV. SCOPE

We are developing a system which will be very helpful for text extraction and recommendation. With rapid developments in text extraction techniques, demand of systems just like text extraction and recommendation is increases day by day. So there is need to develop a smart system which can be helpful for text extraction and recommendation.

V. DESIGN METHODOLOGY

A. Model Approach

The proposed system consists of following modules:

1) Module 1: Login and Registration.

The user opens the system manually. Once the user opens the web app, if it is new user they have to do registration first and after validation checking and authentication, user can login to the system.

2) Module 2: Image Capture.

After successful login user will gives input in the form of image. Two ways to give input as image, by opening camera then capture image or by uploading existing image from folder.

3) Module 3: Text Extraction.

Once the application captures the image, it is processed to remove noises after which the texts are sharpened and then texts are extracted image using OCR. OCR Stands for "Optical Character Recognition." While OCR technology was originally designed to recognize printed text from the image, it can be used to recognize and verify handwritten text as well. Apart from these following four basic steps, OCR accuracy can be enhanced through the implementation of application-specific optimizations:

- 1) Obtain image.
- 2) Perform pre-processing on the image.
- 3) Apply algorithm for character recognition.
- 4) Post-processing.

4) Module 4: Provide Recommendations.

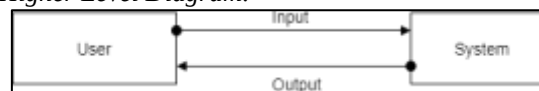
After text extraction, system will provide related recommendations based on that text by using CBDR (content based data recommendation) algorithm. Thus, proposed system will provide recommendations by extracting text from image.

B. Introduction of Design Methodology

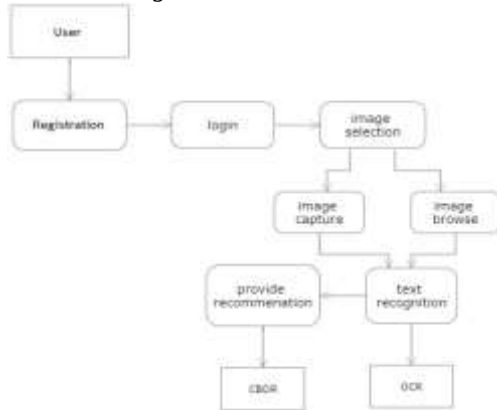
We are developing application for text extraction from images and provide recommendations related to that extracted text to user. The application is very easy to use as it is web application users can use it in there pc. user will capture image or upload the image then by applying OCR optical character recognition method, text will be extracted from image. And after text extraction, by using CBDR content based data recommendation algorithm recommendations will be provided to user.

C. Overall All System Diagram –

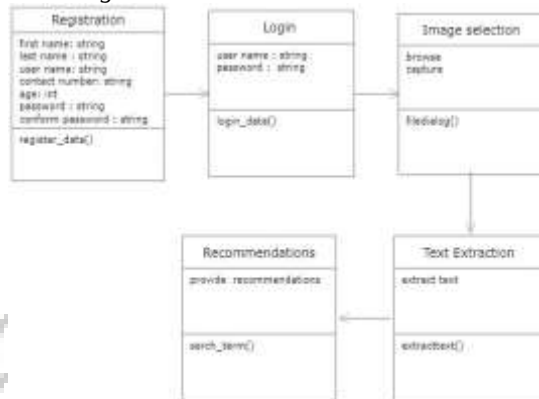
1) Higher Level Diagram:



2) Lower Level Diagram:



3) Class Diagram



VI. FUTURE PROSPECTS

We have developed this system to implement a system that can help the users to dowork in the smart way and in the accurate way. Even this system can work more efficiently as we go further. This system is really useful for the users we can further add extra sections in this software. We can add the crop button also in this software. So proposed system will extract the text from image and provides appropriate recommendations to users. It will saves the time and increases performance of current recommendation system. As we can go further we can also add Text to speech synthesizing for providing recommendations. In this we can extract text from image and convert it to speech because of this speed of the recommendation system is increases and time saves. We can also add multiple languages like Marathi, Hindi, Urdu for text extraction and we can also provide recommendations through multiple languages. This is very useful for the users who don't have the knowledge of English language.

VII. CONCLUSION

Using this system we develop the smart system which extract text from image and provide appropriate recommendations to users. This system is useful reduce there efforts which are require for the text extraction from the images. This system is also provide the recommendations related to the text extracted from the images. This system is very useful for E-commerce. This system is easy to work and saves the time and increases performance of current recommendation system. This system can helpful for users to access anywhere. This system is easy to operate.

VIII. REFERENCES

- [1] Chirag Patel, Atul Patel, Dharmendra Patel,"Optical Character Recognition by Open Source OCR Tool Tesseract: A Case Study", International Journal of Computer Applications (0975 – 8887), Volume 55– No.10, 2012.
- [2] S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, "Data Preprocessing for Supervised Learning", INTERNATIONAL JOURNAL OF COMPUTER SCIENCE VOLUME 1 NUMBER 1 2006 ISSN 1306-4428.
- [3] Sandip Rakshit ,Amitava Kundu,"Recognition of handwritten Roman Numerals using Tesseract open source OCR engine", Proc. Int. Conf. on Advances in Computer Vision and Information Technology (2009) 572-577.
- [4] An Automatic System To Detect And Recognize Text In Images Victor Wu, R. Manmatha, Edward M. Riseman.
- [5] "Journal of Information & Communication Technology- JICT Vol. 10 Issue. 2, December 2016 A Survey on Optical Character Recognition System" Noman Islam, Zeeshan Islam, Nazia Noor.
- [6] "Content-Based Image Retrieval", Masooma Zahra.