ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 9, Issue 11 , November 2022

Extraction of Text from Digital Images in Multiple Languages using Pytesseract

Dr.G.Lakshmi, V.Pravallika, R.Rattaiah, M.Sriramchandra, M.Supritha, T.Lohith

Assistant Professor, Department of Information Technology, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India
B.Tech Student, Department of Information Technology, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India
B.Tech Student, Department of Information Technology, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India
B.Tech Student, Department of Information Technology, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India
B.Tech Student, Department of Information Technology, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India
B.Tech Student, Department of Information Technology, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India

ABSTRACT: The process of data digitalization involves categorising the optical patterns included in a digital image. Segmentation, feature extraction, and classification are used to recognise characters. The fundamental OCR concepts presented in this chapter will help you better grasp the text. An overview of the background and development of OCR systems opens the chapter. Afterwards, the various OCR system methodologies, including optical scanning, location segmentation, pre-processing, segmentation, representation, feature extraction, training and recognition, and post-processing, will be discussed. The various applications of OCR systems are then emphasised, followed by a discussion of the OCR systems' present state. The OCR systems' future is finally presented.

KEYWORDS: Optical Character Recognition, Tesseract, Text Extraction.

I. INTRODUCTION

As we all know, there are numerous printed newspapers and books on a variety of themes. The issue here is for software systems to recognise characters in computer systems when information is scanned through paper documents. When we scan papers with a scanner, the documents are saved in the computer as images in formats like JPEG and GIF. The user cannot read or change these photos. However, it is highly challenging to read the individual contents and search the contents of these documents line-by-line and word-by-word in order to reuse this information. There is a great deal of interest these days in "saving the information present in these paper documents in to a computer storage disc and then modifying or reusing this information via searching.

II.PROPOSED SYSTEM

Our suggested solution is a character recognition system called OCR on a grid infrastructure, which can recognise characters from various languages. This function, which we refer to as grid infrastructure, solves the issue of heterogeneous character recognition and enables a variety of functionality to be applied to the document. While the current system just provides document editing, the multiple functionalities also support editing and searching. In this context, "Grid infrastructure" refers to the system that supports a certain group of languages. OCR is multilingual on a grid infrastructure because of this.

III. TECHNOLOGIES USED

A. Jupyter Notebook

A web-based interactive computational environment for authoring notebook documents is Jupyter Notebook (formerly Ipython Notebook). Several open-source libraries, including IPython, ZeroMQ, Tornado, jQuery, Bootstrap, and MathJax, are used in the construction of Jupyter Notebook. An application called a Jupyter Notebook is a browser-based REPL that has an ordered list of input/output cells that can contain code, text (written in Markdown), math, graphs, and rich media.A Jupyter Notebook document is a JSON file that adheres to a versioned format and typically ends in ".ipynb." The metadata, notebook format, and cell list are the three essential components of Jupyter notebooks. To set up and show the notebook, metadata acts as a data dictionary of definitions. The software's version number is Notebook Format. several sorts of cells are listed. Markdown cells for display, code cells for execution, and code cells for output.Even though ".ipynb" and JSON are the most typical and default formats, it is possible to forego some capabilities (such storing photos and metadata) and save notebook as markdown documents using an addon like JupyText. Version control is frequently used in conjunction with Jupytext to streamline notebook merging and diffing.

B. Python Modules

The Python modules utilised in this instance include OS, opencv2, PIL, and Pytesseract. Python's OS module offers tools for communicating with the operating system. OS is included in the basic utility modules for Python. A portable method of exploiting operating system-specific functionality is offered by this module. There are numerous functions to deal with the file system in the *os* and *os.path* modules.

A sizable open-source library for image processing, machine learning, and computer vision is called OpenCV. Python, C++, Java, and many other programming languages are supported by OpenCV. It can analyse pictures and movies to find faces, objects, and even human handwriting. This OpenCV will assist you in learning image processing, including operations on images and videos, from the very beginning to the more advanced levels.

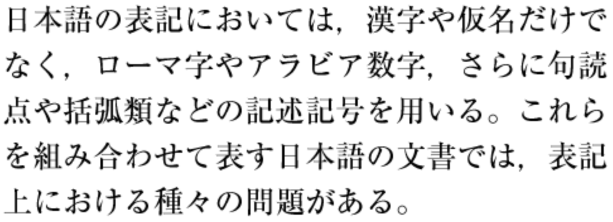
The standard image processing package for the Python language is the Python Imaging Library (extension of PIL). It includes simple image processing capabilities that help with image creation, editing, and saving. An optical character recognition (OCR) tool for Python is called Python-tesseract. In other words, it will identify and "read" any text that is contained in photos.

C. Data

The data is taken in the form of digital images. The images of different languages are collected like Portuguese, English, Hindi, Japanese etc. Using this images and pytesseract tool we can extract the text embedded in images.

IV. RESULTS

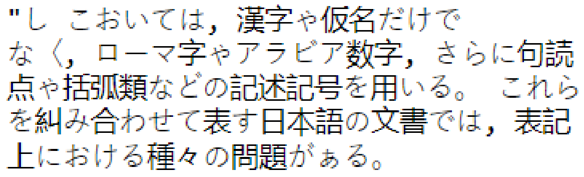
INPUT:



日本語の表記においては，漢字や仮名だけでなく，ローマ字やアラビア数字，さらに句読点や括弧類などの記述記号を用いる。これらを組み合わせて表す日本語の文書では，表記上における種々の問題がある。

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 9, Issue 11 , November 2022

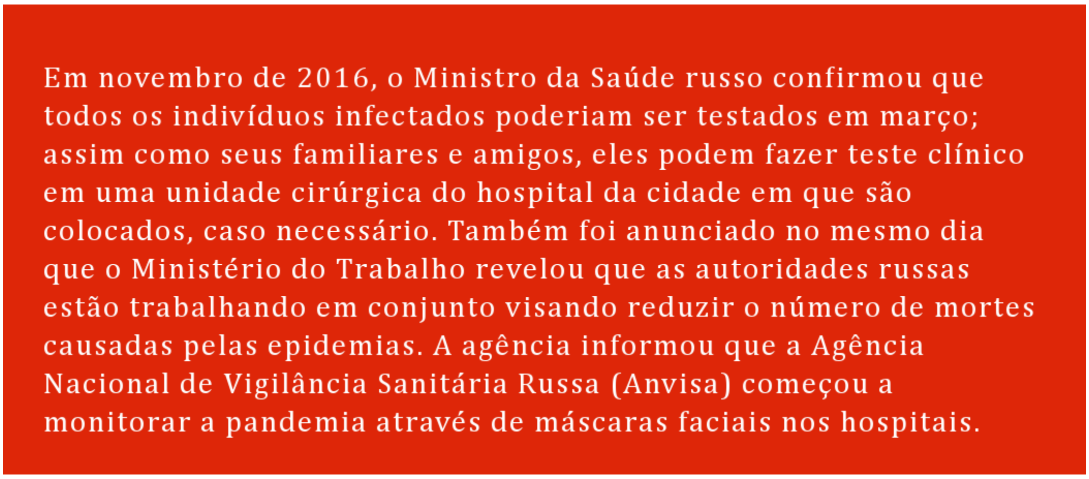
OUPUT:



```
"し こおいては，漢字ゃ仮名だけで
なく，ローマ字ゃアラビア数字，さらに句読
点ゃ括弧類などの記述記号を用いる。 これら
を糾み合わせて表す日本語の文書では，表記
上における種々の問題がぁる。
```

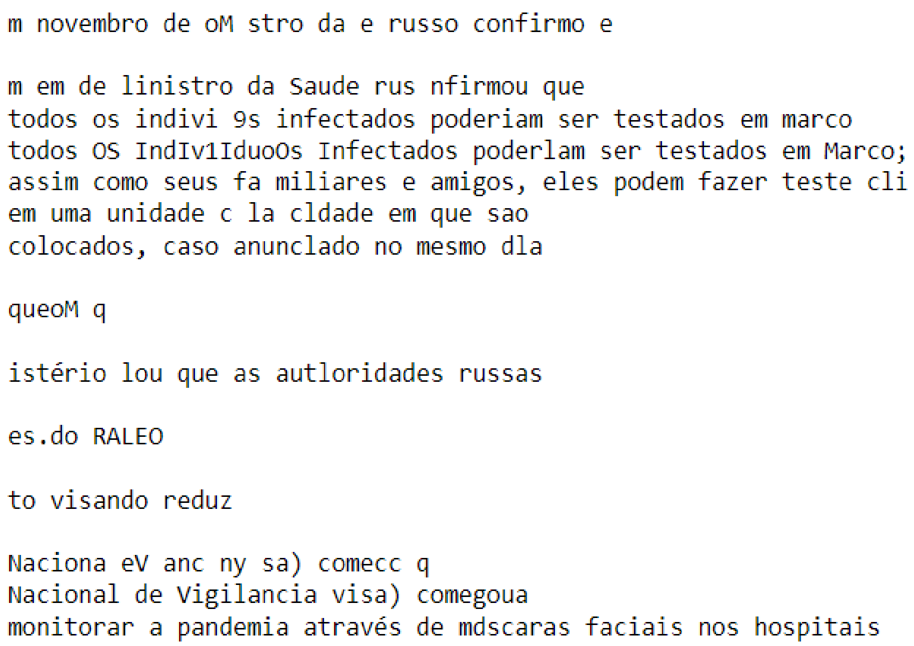
The above image we taken is Japanese text image which we cannot able to copy text from image by using this pytesseract tool the text was extracted from the image which we can edit,modify and can perform any operations on it.

INPUT:



Em novembro de 2016, o Ministro da Saúde russo confirmou que todos os indivíduos infectados poderiam ser testados em março; assim como seus familiares e amigos, eles podem fazer teste clínico em uma unidade cirúrgica do hospital da cidade em que são colocados, caso necessário. Também foi anunciado no mesmo dia que o Ministério do Trabalho revelou que as autoridades russas estão trabalhando em conjunto visando reduzir o número de mortes causadas pelas epidemias. A agência informou que a Agência Nacional de Vigilância Sanitária Russa (Anvisa) começou a monitorar a pandemia através de máscaras faciais nos hospitais.

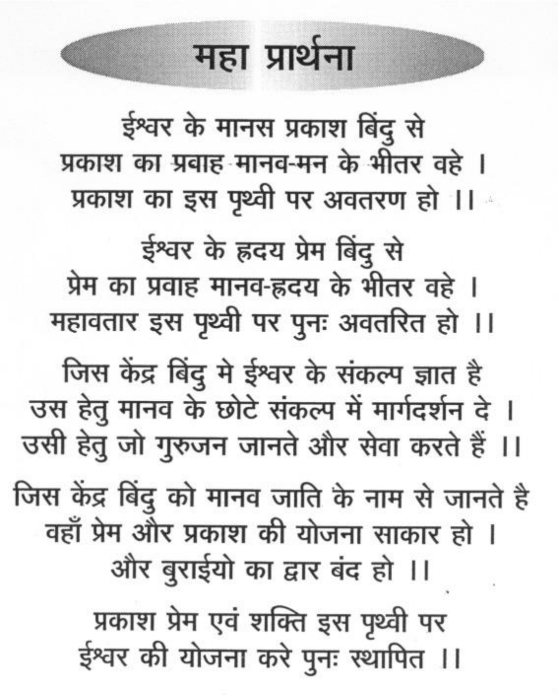
OUTPUT:



```
m novembro de oM stro da e russo confirmo e

m em de linistro da Saude rus nfirmou que
todos os indivi 9s infectados poderiam ser testados em marco
todos OS IndIv1IduoOs Infectados poderlam ser testados em Marco;
assim como seus fa miliares e amigos, eles podem fazer teste cli
em uma unidade c la cldade em que sao
colocados, caso anunclado no mesmo dla

queoM q

istério lou que as autloridades russas

es.do RALEO

to visando reduz

Naciona eV anc ny sa) comecc q
Nacional de Vigilancia visa) comegoua
monitorar a pandemia através de mdscaras faciais nos hospitais
```

The above image we taken is Portuguese text image which we cannot able to copy text from image by using this pytesseract tool the text was extracted from the image which we can edit,modify and can perform any operations on it.

INPUT:



महा प्रार्थना

ईश्वर के मानस प्रकाश बिंदु से
प्रकाश का प्रवाह मानव-मन के भीतर वहे ।
प्रकाश का इस पृथ्वी पर अवतरण हो ।।

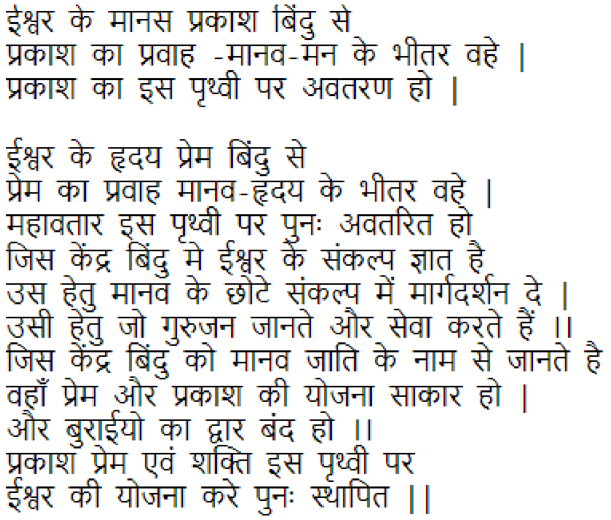
ईश्वर के ह्रदय प्रेम बिंदु से
प्रेम का प्रवाह मानव-ह्रदय के भीतर वहे ।
महावतार इस पृथ्वी पर पुनः अवतरित हो ।।

जिस केंद्र बिंदु मे ईश्वर के संकल्प ज्ञात है
उस हेतु मानव के छोटे संकल्प में मार्गदर्शन दे ।
उसी हेतु जो गुरुजन जानते और सेवा करते हैं ।।

जिस केंद्र बिंदु को मानव जाति के नाम से जानते है
वहाँ प्रेम और प्रकाश की योजना साकार हो ।
और बुराईयो का द्वार बंद हो ।।

प्रकाश प्रेम एवं शक्ति इस पृथ्वी पर
ईश्वर की योजना करे पुनः स्थापित ।।

OUTPUT:



```
इश्वर के मानस प्रकाश बिंदु से
प्रकाश का प्रवाह -मानव-मन के भीतर वहे |
प्रकाश का इस पृथ्वी पर अवतरण हो |

ईश्वर के हृदय प्रेम बिंदु से
प्रेम का प्रवाह मानव-हृदय के भीतर वहे |
महावतार इस पृथ्वी पर पुनः अवतरित हो
जिस केंद्र बिंदु मे ईश्वर के संकल्प ज्ञात है
उस हेतु मानव के छोटे संकल्प में मार्गदर्शन दे |
उसी हेतु जो गुरुजन जानते और सेवा करते हैं ||
जिस केंद्र बिंदु को मानव जाति के नाम से जानते है
वहाँ प्रेम और प्रकाश की योजना साकार हो |
और बुराईयो का द्वार बंद हो ||
प्रकाश प्रेम एवं शक्ति इस पृथ्वी पर
ईश्वर की योजना करे पुनः स्थापित ||
```

The above image we taken is Hindi text image which we cannot able to copy text from image by using this pytesseract tool the text was extracted from the image which we can edit,modify and can perform any operations on it.

V. SCOPE FOR FUTURE USE

This has the potential to develop into a potent tool for upcoming data entering applications. Data entry that is automated is one of the most appealing and labour-saving technologies. The technology recognises new font characters very quickly and easily. The document's material can be modified more easily, and the revised content can be reused as needed. Other than editing and searching, software extension is a future research area. By making it more user-friendly, training and recognition speed may be raised steadily.

VI. CONCLUSION

Data Digitalization deals with recognition of optically processed characters. Reliably interpreting text from real-world photos is a challenging problem due to variations in environmental factors even it becomes easier using the best open source OCR engine.

REFERNCES

1)https://moov.ai/en/blog/optical-character-recognition-ocr/

2)https://www.slideshare.net/nikbharat/project-report-of-ocr-recognition

3)https://nanonets.com/blog/ocr-with-tesseract/

4)https://www.analyticsvidhya.com/blog/2020/05/build-your-own-ocr-google-tesseract-opencv/

5)https://en.wikipedia.org/wiki/Optical_character_recognition

6)https://www.youtube.com/watch?v=Rb93uLXiTwA

7) https://web.archive.org/web/20160415060125/https://dev.havenondemand.com/apis/ocrdocument