Journal of Research in Engineering and Applied Sciences

# IMAGE TEXT EXTRACTION AND ITS LANGUAGE TRANSLATION

Shubham Nagmoti[1], Kapil Bhoyar[2], Shantanu Raut[3], Saransh Jamgade[4], Nikhil Mangrulkar[5] and Aniket Pathade[6]

[1-4] Student, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India.

[5]Assistant Professor, Department of Computer Technology, Yeshwantrao Chavan College of Engineering, Nagpur, Maharashtra, India.

[6]Research Consultant, Jawaharlal Nehru Medical College, Datta Meghe Institute of Medical Sciences, Sawangi (M), Wardha, Maharashtra, India.

**Abstract-** Nowadays paperless offices and digitizing document is becoming ordinary for every kind of business or work. It is a good idea to find an easy way to create, store, and protect important documents. Document scanning can be the way Unlike the traditional manual method of creating and preserving document which comes with many benefits such as more office space, information storing, sharing, better data security and recovery. In This paper we have proposed about document scanning in terms of a software interface i.e. web application that does an automated digitization of document with various features such as image enhancement, perspective view, text recognition and translation. Also, we have discussed about various Python scripts and web frameworks used for achieving optimal document scanning.

## 1. Introduction

Document scanning has immensely evolved ever since digital era. It is important step or the first step in text recognition and image enhancement. Perspective transformation is the first step towards text recognition. To get the top down view of a 3D image, we use perspective transformation. It helps to get better insight of the data in images.

Optical character recognition (OCR) is the electronic conversion of images of handwritten or printed text into electronic format. OCR has series of steps which includes Image acquisition, Prepossessing, Segmentation, Feature Extraction and Classification. Tesseract is an optical character recognition module for python. The module is designed to read the text from images in JPEG, GIF, PNG etc. Tesseract works on segmentation by differentiating the background and foreground of the image and adaptive recognition technique by matching pixels of the characters. Storing the content of Books, paper documents, newspapers etc. into the electronic format i.e. computer readable format is the primary task of OCR systems. Later the data in electronic format can be used for various further post processing like language translation, changing fonts etc. OpenCV is a library in python by which we can apply perspective transformation images. There are several factors which defines the quality of image. These factors mainly include the noise, blur, uneven lighting, distortion, contrast, resolution etc. For overcoming the shortcomings created by the factors image enhancement is used. If any of these factors is found significantly affects the text recognition so the is important. This paper is divided into 3 sections – section 2 includes introduction, section 3 presents the survey work related to field of document scanning and text recognition, section 4 discusses about the implementation of project.

## 2. Literature survey

Pooja Sharma *et. al.* has presented a survey work that has been performed in the field of document scanning. Some other features include quality assessment methods and metrices for document images [1]. Also, there are some work done related to the web application translation of text in any language with help of python language libraries and Google Translator is good example of such topics. There are various steps in this which include pre-processing of document, character recognition, segmentation, text extraction, detection and translation of language. These factors when evaluated properly results in image enhancement and better text recognition [2]. There are some problems in text recognition. OCR being an evolving technology, is not 100% accurate while recognizing text and human inspection is necessary. Zeev Zelavsky *et. al.* suggested an algorithm for recognizing text based on fuzzy logic depending on data of the font. This procedure proposes a way for recognition of distorted letters using statistics and fuzzy logic. Their focus was recognition of text of Bible written in calligraphy [3]. Another such work is done by Badawy *et. al.* on Automatic license plate recognition (ALPR) is related to text recognition which extracts the number from the number plate and the information about the vehicle. The information extracted can be used in many applications, such as toll n payment, parking fee payment, and freeway and arterial monitoring systems for traffic surveillance. The ALPR uses infrared camera to take images [4]. Recognition of text from document images as a process of an Optical Character

Recognition (OCR) system is important concept [5]. There are different technologies for automatic identification and establishing position of OCR among these techniques. Shyam G.Dafe *et. al.* have presented different steps that are involved in the OCR system[6].

### 3. Implementation

This project is implemented in python using its image processing library OpenCV and tesseract. For language translation google translation API (googletrans) is used.

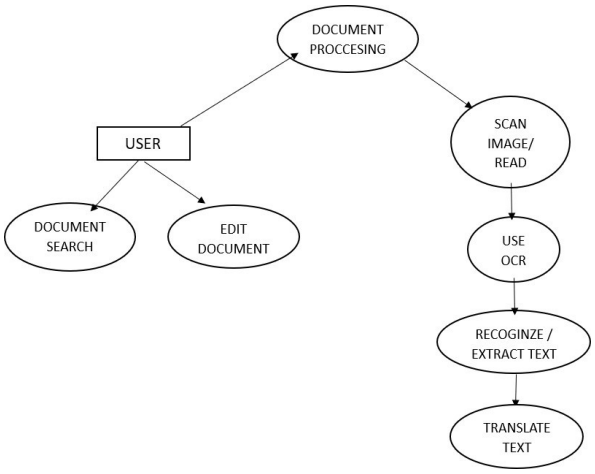



Fig. 1: Block diagram of OCR system

The above Fig. 1 shows block diagram is of the OCR system which gives us the brief idea about the processing. Firstly, there is scanning(reading) task which is the input to the system. Then the OCR methodology is applied to that read image. Using tesseract module text is extracted from the input image. This extracted text can be then proceeded to the translation into different languages.

1)Perspective transformation:

Perspective transformation is the first step towards text recognition. To get the the top down view of a 3D image, we use perspective transformation. It helps to get better insight of the data in images. OpenCV is a library in python by which we can apply perspective transformation images.

2) Text recognition:

Tesseract is an optical character recognition module for python. The module is designed to read the text from images in JPEG, GIF, PNG, etc. Tesseract works on segmentation by differentiating the background and foreground of the image and adaptive recognition technique by matching pixels of the characters.

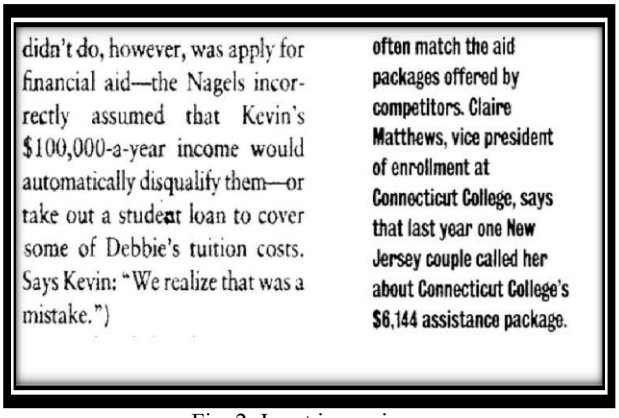
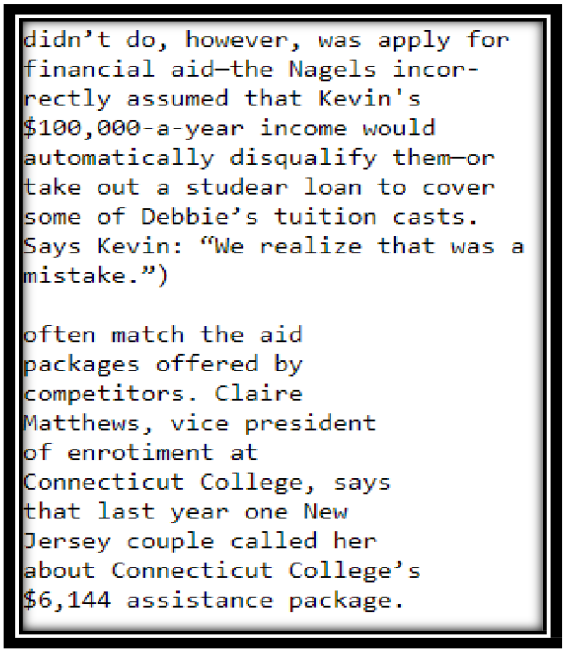



Fig. 2: Input image.jpg

Fig. 2 is text image as the input to the OCR system and it goes through two passes of recognition in tesseract. Adaptive thresholding is done on the




input image to make it a binary image. Now the system checks for lines and words in the image. After the two passes of recognition, the text is extracted

Fig 3: Generated output

The figure (Fig. 3) shows the extracted text from the input image using tesseract module that is imported in python code.

3)Language Translation:-

Google translate API can be used to translate a text into other languages known to Google translator. Googletrans is a free python library that implemented google translate API.

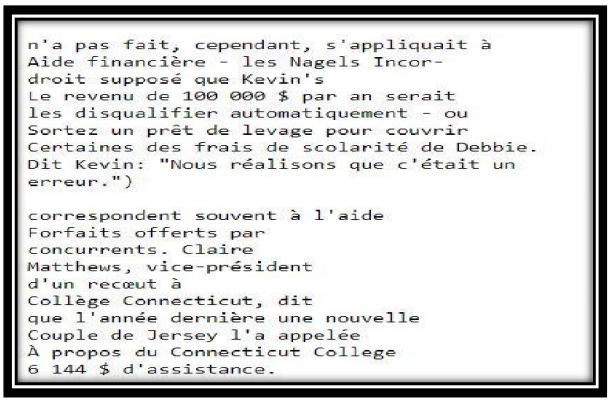



Fig 4: Text translated to French language

Fig. 4 is output that we get in previous step is translated into the French language using the google translate API that is imported in python code .It is an open source , free to use multilingual neural machine translation service developed by Google with each language having its own abbreviation.

## 1. Conclusion

In this paper, we have studied and discussed different methodologies for image acquisition to overcome challenges like a blur, low resolution, uneven lightening conditions, etc. We have also reviewed some methods and algorithms for character recognition from the document images. It can be used to convert the document images into characters which can be stored as electronic format readable by computers and then can be translated into any language known to Google Translate API. By reviewing and analysing such algorithms and methodologies we deduced that any document images available with us can be read and can be converted into an electronic format and is ready for post processing like language translation, changing fonts etc.

## Reference

[1] P. Sharma and S. Sharma, "An analysis of vision based techniques for quality assessment and enhancement of camera captured document images," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 425-428, doi: 10.1109/CONFLUENCE.2016.7508157.

[2] S. Thakare, A. Kamble, V. Thengne and U. R. Kamble, "Document Segmentation and Language Translation Using Tesseract-OCR," 2018 IEEE 13th International Conference on Industrial and Information Systems (ICIIS), Rupnagar, India, 2018, pp. 148-151, doi: 10.1109/ICIINFS.2018.8721372.

[3] E. Gur and Z. Zelavsky, "Retrieval of Rashi Semi-cursive Handwriting via Fuzzy Logic," 2012 International Conference on Frontiers in Handwriting Recognition, Bari, 2012, pp. 354-359, doi: 10.1109/ICFHR.2012.262.

[4] S. Du, M. Ibrahim, M. Shehata and W. Badawy, "Automatic License Plate Recognition (ALPR): A State-of-the-Art Review," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 2, pp. 311-325, Feb. 2013, doi: 10.1109/TCSVT.2012.2203741.

[5] S. Malakar, S. Halder, R. Sarkar, N. Das, S. Basu and M. Nasipuri, "Text line extraction from handwritten document pages using spiral run length smearing algorithm," 2012 International Conference on Communications, Devices and Intelligent Systems (CODIS), Kolkata, 2012, pp. 616- 619, doi: 10.1109/CODIS.2012.6422278.

[6] Shyam G.Dafe, Shubham S. Chavhan , “Optical Character Recognition Using Image Processing”, International Research Journal of Engineering and Technology (IRJET)e-ISSN: 2395-0056 Volume: 05 Issue: 03 | Mar-2018.