Techorrect: A Tool To Evaluate Answer Scripts

Divya Kumari Tankala, R Mamatha, P Neelima
Department of Computer Science & Engineering
G Narayanamma Institute of Technology and Science for women, Hyderabad,Telangana,India

Abstract- Evaluate hundreds of papers is a very tedious and time consuming process. The Evaluator needs to have the knowledge of the subject to correct papers without which one cannot correct them. Absence of evaluator makes the process even longer. We are moving towards the digital world where every problem can be solved with a laptop in our lap. With that intention and using technologically advanced systems, this paper gives a solution, is to correct objective papers using character recognition technique from the concepts of machine learning instead of doing it manually.

Keywords – Statistical techniques of machine learning, binarization and pre-processing images

I. INTRODUCTION

Now a days, Evaluating answer scripts is tedious task. The evaluator need to have knowledge of the subject to correct papers without which no one can correct them. Using this even if the evaluator doesn't have any knowledge about the subject, papers can be corrected and marks can be allotted. Time to evaluate papers is also reduced. Hundreds of papers can be evaluated in a short span of time and correction can be done from anywhere. In this paper, evaluating objective paper using Optical Character Recognition technique(OCR) from the concept of machine learning instead of doing it manually. To make the process of tests and assessment, evaluation and feedbacks, easier which can be used in schools, institutions, national and international exams, colleges, any competitive exam and integrating it into daily lives to help the students and the faculty.

II. METHODOLOGY

Today, the education system is totally changed by the technology. It has become more interesting and informative by projector teaching, online tutorials, online teaching video and animation etc. There is an intense use of technology to teach the student. But evaluation process is still done in the traditional way. Objective test demands one specific answer from multiple answers. It can measure all levels of student's ability from memory to synthesis. It enables wide sampling of subject content, quick and easy to score. This type of test is used to examine the students frequently due to quite economical. Objective test can be conducted online and offline modes. In offline method we generally use Optical Mark Recognition / Optical Mark Reading(OMR) sheet. In first sight, online mode seems good in comparison to offline mode without any paper cost, automatic and fast evaluation. Manual evaluation of offline objective answer-sheet is time consuming process because at the time of evaluating answer, we must evaluate it from the ideal sheet answer-sheet if there are 60 questions then we must match 60 times with ideal sheet and mark it right or wrong. In this paper, proposed a tool which scan the objective paper and pass it onto the algorithm which will convert letters in the image to text, after conversion the correct options which are already stored in the database created and maintained by the operator, will be checked against the options written by the student. The role of operator is to upload the student and faculty details to the database. The final marks after the evaluation will be displayed in a system that is accessible by both students and faculty. The students can login and check their marks with respect to subject.

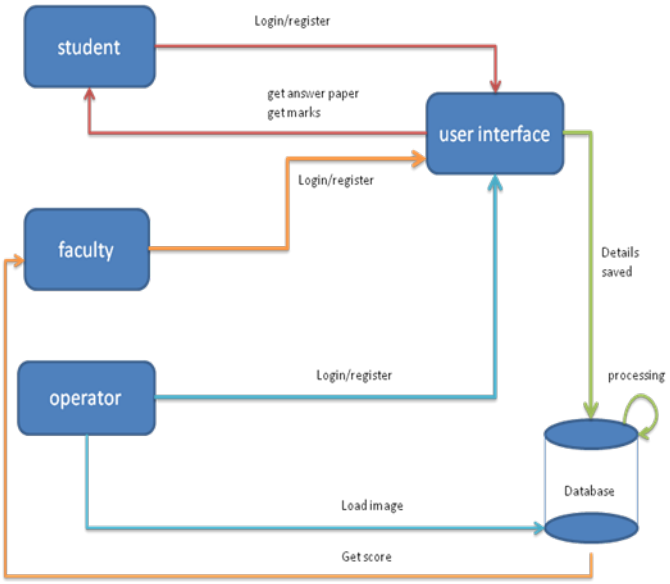

Figure 1: Architecture of system

2.1 OCR: Optical character recognition:

OCR is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo (for example the text on signs and billboards in a landscape photo) or from subtitle text superimposed on an image (for example from a television broadcast). Widely used as a form of information entry from printed paper data records whether passport documents, invoices, bank statements, computerized receipts, business cards, mail, printouts of static-data, or any suitable documentation it is a common method of digitizing printed texts so that they can be electronically edited, searched, stored more compactly, displayed on-line, and used in machine processes such as cognitive computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision. The OCR algorithm relies on a set of learned characters. It compares the characters in the scanned image file to the characters in this learned set. Generating the learned set is quite simple. Learned set requires an image file with the desired characters in the desired font be created, and a text file representing the characters in this image file.

Hand writing character recognition is a very tough job due to different writing style of user as well as different pen movements by the user for the same character.

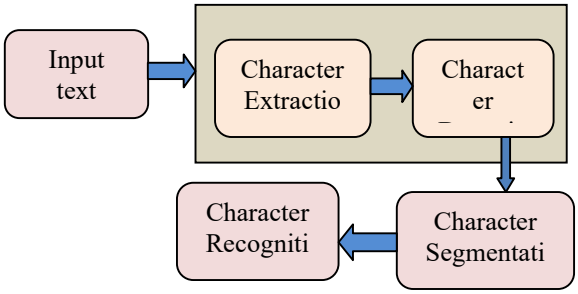

Figure 2: Methodology block diagram

The process of OCR is a composite activity comprises different phases as follows:

Image acquisition: To capture the image from an external source like scanner or a camera etc. the approaches can be

Digitization, binarization and compression

Preprocessing: Once the image has been acquired,

Various preprocessing techniques can be performed to improve the quality of image. Such techniques are noise removal, thresholding and extraction image base line etc.
Character segmentation: the characters in the image are separated such that they can be passed to recognition engine. Among the simplest techniques are connected component analysis and projection profiles can be used.
Feature Extraction: The segmented characters are then processes to extract different features. Such features should be efficiently computable, minimize intra-class variations and maximizes inter-class variations.
Character classification: It maps the features of segmented image to different categories of classes. Structural classification techniques are based on features extracted from structure of image and uses different decision rules to classify characters. Statistical pattern classification methods are based on probabilistic models. The techniques used are Neural network, Bayesian, Nearest Neighborhood etc.
Post processing: After classification, the results are not 100% correct, especially for complex languages. Post processing techniques can be performed to improve the accuracy of OCR Systems.

III. IMPLEMENTATION

The sample code shows how handwritten text is recognized for a given question paper:

```
//open image
img=Image.open('C:\\img\\14-63_LP.jpeg')
//Crop the specified area
img1=img.crop((2179,844,2284,907))
//Save it in another file
img1.save('C:\\Users\\Downloads\\img1.jpg')
//Extracting characters from the image using pytesseract algorithm
x=get_string("C:\\Users\\Downloads\\img1.jpg")
```

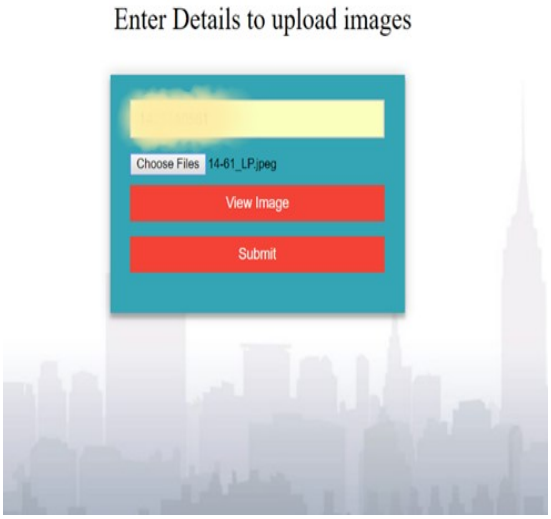

Figure 3:(a) sample code to upload paper (b) Screen shot to upload answer script

Tesseract algorithm can accurately decipher and extract text from a variety of sources! As per its namesake it uses an updated version of the tesseract open source OCR tool. We also automatically binarize and pre-process images using the binarization, so tesseract has an easier time deciphering images. Not only are we able to extract English text, but tesseract supports over 100 other languages as well. This algorithm works best when the image only contains text and is on a single line. It's recommended to crop out everything else from the image with an algorithm like text detection first. It is an optical character recognition engine for various operating systems. It is free software, released under the Apache License, Version 2.0 and development has been sponsored by Google since 2006. Tesseract is suitable for use as a backend and can be used for more complicated OCR tasks including layout analysis by using a frontend.
Pytesseract Pytesseract (Python-Tesseract) is an OCR tool for python. That is, it will recognize and “read” the text embedded in images. Pytesseract is a wrapper for Google's Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Python Imaging Library, including jpeg,

png, gif, bmp, tiff, and others, whereas tesseract-ocr by default only supports tiff and bmp. Additionally, if used as a script, Python-tesseract will print the recognized text instead of writing it to a file.

IV.CONCLUSION

In this paper, The system is viable only when the bracket area where the option has to be written is fixed. The project can be further enhanced by detecting the hand writing in blanks too and giving the marks on a whole. This is the faster way to evaluate offline objective answer-sheets. That is really made fast and efficient the evaluation process in education system. The efficiency of system is still required to improve with take care of other issue of image processing like poor scanning, skew correction, poor writing etc. The options of questions may be in other languages like English, Hindi and Punjabi etc. we can make this system work with other language also.

REFERENCES

[1] https://www.pyimagesearch.com/2017/07/10/using-tesseract-ocr-python/

[2] https://pythonhosted.org/spyder/

[3] Ray Smith Google Inc., “An Overview of the Tesseract OCR Engine” in ICDAR 2007

[4] R.W. Smith, the Extraction and Recognition of Text from Multimedia Document Images, PhD Thesis, University of Bristol, November 1987.

[5] Lund, W. B., Kennard, D.J., & Ringger, E.K(2013). Combining Multiple Thresholding Binarization values to improve OCR output presented in Document Recognition and retrieval XX conference 2013, California, USA,2013.