DIGITIZATION OF DOCUMENTS USING OCR

Disha Munde[1], Swarali Kulkarni[2], Anagha Daphal[3]

[1-3]UG Students, Department of Computer Engineering, Marathwada Mitra Mandal's College of Engineering, SPPU, Pune, India.

---***---

Abstract - Nowadays evaluating exam papers and declaring the result in a stipulated period of time is a difficult job for educational colleges, institutions, departments and Universities. Thus physical exam paper evaluation becomes a tedious task. To make it easier and more correct the proposed aim is to develop a software for automatic exam paper evaluation and grading system, the system works by scanning the handwritten written exam papers then the scanning the image be improved into an editable text using OCR tool and the evaluation will perform by matching the key terms which is maintained in the database. It is an entirely integrated approach upon dissimilar levels of knowledge by the method of examination, evaluation, result and formulation of subject papers. In a field of education through teaching, evaluation and the performance method many organizations initiated with the use of technologies. The system works in different patterns such as online and the manual exam paper evaluation with different analysis and techniques.

Key Words: Optical Character Recognition, Machine Learning, Evaluation System, Pre-processing Feature Extraction, Tesseract, NLP

1. INTRODUCTION

In the current evaluation system, descriptive answers given by the students are evaluated manually by the teachers. This is a tedious task, as different teachers are likely to award different marks to the same answer. Teachers have to put a lot of manual effort to read through and evaluate the handwritten papers of all students. When a human being evaluates anything, the quality of evaluation may vary along with the emotions of the person. Performing evaluation through computers using intelligent technique ensures uniformity in marking as the same inference mechanism is used for all the students. This software application provides an automatic exam paper evaluation and grading system for various educational school, colleges ,universities etc. Application also provides automatic result generation. The process can be monitored through various dashboard by admin/senior.

In this system, the handwritten examination papers are scanned and converted into an editable format using OCR tool and then evaluation will be performed by formulas, in-between steps and final solution for the exam papers, for the theory papers the evaluation will be performed by the keywords or synonym based keywords which are maintained in the database

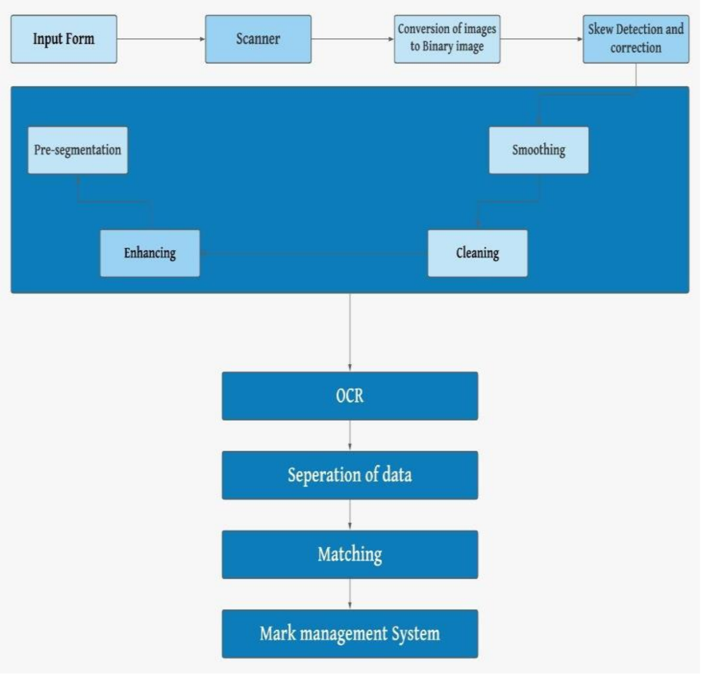
RELATED WORK

The system for identifying text though handwritten answer sheets and also evaluate marks intended for every small answer on the base of acquired knowledge besides a model. OCR tool is used to take out the handwritten text, where NLP and neural networks are used for extraction of keywords though human evaluation dataset samples of handwritten answer paper and keys[2].The system also evaluates The score is based upon similarity measures of sentences. The evaluation method to calculate the score for every answer which is written by the students, anywhere the appliance is trained based upon datasets[3].

The Handwritten Short Answer Analysis System is an automatic small answer evaluation system that is capable of identifying the text in the handwritten answer sheets and evaluating marks for every small answer supported by earlier information acquired besides models. Within the system, Optical Character Recognition tools are accustomed to extract the written texts. Natural language process is employed to take out keywords as a human evaluated model dataset of written answer key and answer paper. The projected models evaluate scores supported circular function sentence similarity measure

1. Proposed System

To overcome the limitations of existing system, the proposed system aim is to build software for evaluating the handwritten examination papers and the theoretical papers automatically. In this system, the handwritten papers are scanned and converted into an editable format using OCR tool and then evaluation will be performed by formulas, in-between steps, the theory papers the evaluation will be performed by the keywords or synonym based keywords which are maintained in the database. Finally the marks will be given depending upon the identical key terms.



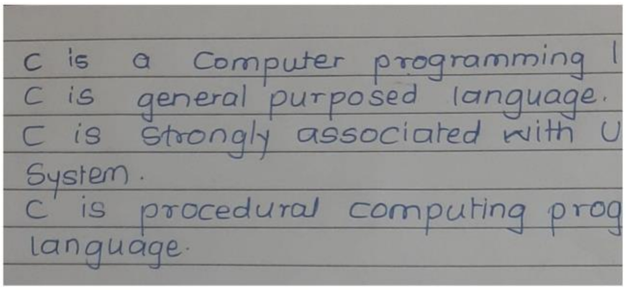
1.1 Architecture Design

A. OCR Tool:

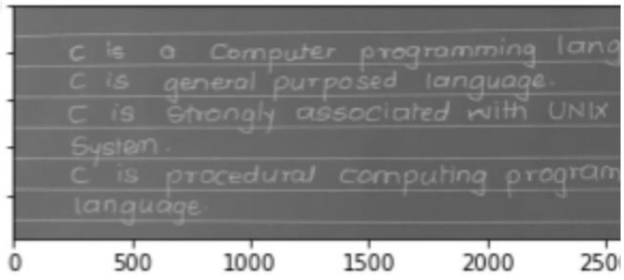
Optical Character Recognition tool that permits to convert the dissimilar types of scanned paper documents, captured images by digital camera upon an editable form or PDF files In order to re-purpose and extract data from camera images, scanned documents or PDFs, the OCR software would particular exposed letters on image and locates them into words then words to sentences by enable to edit and access the original document content[4].

B. Binarization:

Binarization converts a picture from grayscale or color keen on black along with white known as a ”binary image” as present is two colors. It also converts the acquired form of image into binary format, inside which the foreground contains symbols, filled data, frame line and printed entities.



C is a Computer programming l
C is general purposed language.
C is strongly associated with U
System.
C is procedural computing prog
language.



C is a Computer programming lang
C is general purposed language.
C is strongly associated with UNIX
System.
C is procedural computing program
language.

0
500
1000
1500
2000
250

C. Scanning:

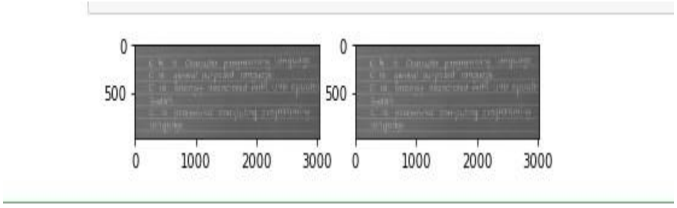
Scanning Towards the model selection for the system process a figure recognition step matches the features which are extracted from an handwritten scanned paper next to where the extraction from each design of a module.

D. Image extraction :

Image extraction The extraction from relevant data though particular fields and preprocesses the information in order to enhance data and eliminate noise. The characters are recognized by the extraction of field is nearby subsequent to
the present analysis.

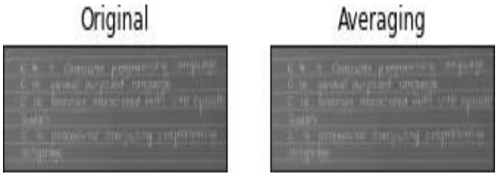
E. Pre-processing Stage

Pre-processing step involves RGB to grayscale image alteration, noise removal, Thresholding etc. In RGB to grayscale picture conversion, RGB image is converted to gray image. Thresholding converts gray to black and white images. If needed the key document could be resized if needed. This Image pre-processing step has an immense impact on the quality of the OCR process.



F.Smoothing :

Smoothing operations are used for noise decrease and for blurring in color images. Blurring used inside the preprocessing stage for removal of undersized details from the image. Smoothing operations reduce the noise in binary
image.



G.Enhancing:

It reconstructs the stroke that has been detached throughout cleaning process. Over the baseline if some filled data is printed then in cleaning it also remove the foremost discontinuity in character which might obstruct the recognition.

H. Character Recognition:

Recognition is to map the given pattern with internally stored database. The matching through the recognition process is presented by the gesture with the standard indication. The important features like the size, depth, shape, color are extracted from the recognition system by input image. If the geometrical features are extracted as input after that the character is able to be coordinated by means of normal characters within the library. Thinning process will remove the pixels consequently the object with no holes shrink toward a minimally associated stroke. Skeletons are useful for describing properties of a shape and also in various cases it is intended for transformation of the original shape.

2. Methodology

2.1 Data Collection

Dataset will be responsible for converting answers into score based on its relevance as per data collected
. We are taking a number of questions, our dataset has 6 questions and every question is graded based on the set of parameters and features described as important by the human grader.

2.2 Data Preprocessing

The dataset contains scanned copies of hand written answer sheets of students. Now, the data preprocessing is done on image by using various processing techniques.such as Binarization, Skew Detection and Correction, Imageextraction, Smoothing, Enhancing, Cleaning and Pre-Segmentation.

2.3 Optical Character Recognition(OCR)

After data preprocessing we have done OCR on that preprocessed image. For this we have used the Py Tesseract tool. It recognizes the text as
black-on- white text, organization of blobs into text lines, breaking text lines into words according to the character spacing kinds and so on[5].

Rather than using raw counts, we are using term frequency – inverse document frequencies.

$$tfidf = (1 + log(tf)) \cdot log\left(\frac{N}{dfw}\right)$$

2.4 Modeling

The proposed system is using the nltk library.

We have used NLP for matching text objects to find similarities[6]. Important applications of text matching include automatic spelling correction, data deduplication, key matching and grammar check etc. In nltk, we used the wordnet library for identification of similar meaning words or synonyms.

2.5 Prediction

The prediction of score is done by calculating average marks in all the processes. And the final score will be displayed on the system.

CONCLUSIONS

Use of computers have considerably increased in educational institutes or organizations. Automated evaluation does not just implement the preset examination and assessment process although it provides reliable and quick appraisal with more flexibility. The important concept is to minimize the manual work and convert all forms of documents into digital form. The person with the minimum knowledge of computers can also use this system easily.

Some extensions may be data extensions, in other words more data can be stored like more subjective type answers or objective type so it would generate more accurate results. We can add various features like sending automated messages regarding results of every student. Students can also request for retest if students fail.

REFERENCES

[1] Shivam Tomar Shubham Sharma N.K. Bansode Prayag Singh, Saurabh Sheorain. “descriptive answer evaluation”. 2018.
[2] Surekha Mariam Varghese Sijimol. P.J. “short answer scoring system using neural networks”. In International Journal of Computer Mathematical Sciences IJCMS ISSN 2347 – 8527 Volume 7,, 2018.
[3] Sijimol. P.J, Surekha Mariam Varghese, “Short Answer Scoring System Using Neural Networks”, International Journal of Computer & Mathematical Sciences IJCMS ISSN 2347 – 8527 Volume 7, April 2018
[4] Abin M Sabu, Anto Sahaya Das. A Survey on various Optical Character Recognition Techniques, IEEE Conference on Emerging Devices and Smart Systems ICEDSS 2018.
[5] Chandni Kaundilya, Diksha Chawla, Yatin Chopra. Automated Text Extraction from Images using OCR System,2019.
[6] Prayag Singh, Saurabh Sheorain, Shivam Tomar, Shubham Sharma, N.K. Bansode, “Descriptive Answer Evaluation” , International Research Journal of Engineering and Technology (IRJET) Volume: 05 ,May 2018
[7]G.Chandralekha, K.Iswarya, M.pajany, P.Vimala “A Survey On Smart Exam Script Evaluation Using OCR And Ontology”, International Journal of Pure and Applied Mathematics, Volume 119 No. 14, 2018.