BANGALORE INSTITUTE OF TECHNOLOGY

K.R. Road, V.V.Pura, Bengaluru-560004

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Project Work Synopsis

VII-Sem 2019-2020

PROJECT GROUP:

Sl. No.	USN	Name	Section	Phone Number	Signature
1	1BI16CS158	SUJAY KUMAR P S	C	9945562617	
2	1BI16CS173	VADIRAJA RAO M K	C	9482447043	
3	1BI16CS181	VISHVESHWARA G	C	7760402418	
4	1BI16CS187	LOKESHWAR S	C	9538995630	

PROJECT DETAILS:

Title:	**Genrating searchable database of analog documents using optical character recognition and keyphrase extraction.**
Domain:	**Machine Learning**
Location:	**Bengaluru**

For office use only:

Group ID:	
Guide:	
Status:	**Accepted/To be modified/Rejected**

Signature of the Project Co-Ordinator

TITLE

Genrating searchable database of analog documents using optical character recognition and keyphrase extraction.

LITERATURE SURVEY

To extract characters that belong to a document image we make use of Optical Character Recognition (OCR). The universal OCR system consists of three main steps which are image acquisition and preprocessing, feature extraction and classification. Image preprocessing phase cleans up and enhances the image by noise removal, correction, binarization, dilation, color adjustment and text segmentation. Feature extraction is to extract and capture information from the acquired text image to be used for classification. In the classification phase, the portion of the segmented text in the document image is mapped to the equivalent textual representation.

Keyphrases of a given document represent its main topic and they are used as a simple method to represent the document. Using any single method of keyphrase extraction will not provide accurate keyphrases of any given document, due to the design of the algorithm which inherently fails to account for few cases. To improve the accuracy and quality of keyphrase extraction, we use multiple keyphrase extraction algorithms like TextRank, PF-IDF, Graph-based, statistical method etc.

There is no existing system that provides a searchable database of printed document images, which has to make use of OCR and keyphrase extraction.

PROBLEM DEFINITION

There is an abundant information that reside in analog form, which makes it extremely strenuous to search through the documents by hand since it cannot be done using a machine. This problem can be overcome by digitizing the analog documents in a very simple form, like using a scanner or a camera, and then converting it into text format, using OCR, that the machine can handle easily. Then this ocean of data can be made searchable using keyphrase extraction techniques on the text version of the document.

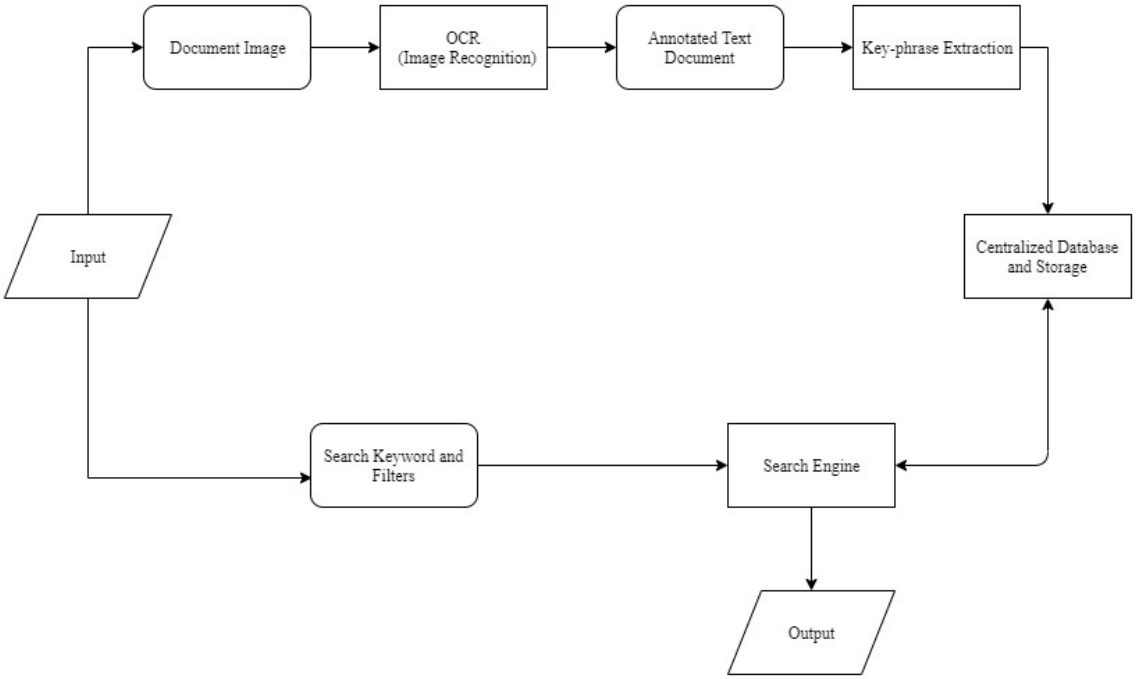
PROPOSED SYSTEM

The analog document is given in the form of a set of images or a single image. Then we apply an OCR algorithm to convert the image into an annotated text format. We apply a keyphrase extraction technique to extract important keywords in the given document to make it searchable and summarizable.We later store this annotated text along with the document images and keyphrases in a database, to provide users an ability to search for revelant documents of a given topic in realtime.

SYSTEM REQUIREMENTS

- Python 3.x
- Linux operating system
- Editors (VSCode, Emacs, Sublime)
- Version control system (git)
- Remote repository (Github)
- Database software (MySQL, Firebase Realtime Cloud)
- CPU Intel 2GHz
- 4GB RAM
- 100GB persistent storage

SYSTEM ARCHITECTURE



APPLICATIONS

- Digitizing of records that are hardcopy materials and fragile in nature.
- Digitally searching materials for a particular topic.
- Copying Text from the analog form to reusable digital format.

REFERENCES

[1] Wei, T. C., Sheikh, U. U., & Rahman, A. A.-H. A. (2018). *Improved optical character recognition with deep neural network. 2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA).*

[2] Yeom, H., Ko, Y., & Seo, J. (2019). *Unsupervised-learning-based Keyphrase Extraction from a Single Document by the Effective Combination of the Graph-based Model and the Modified C-value Method. Computer Speech & Language.*