Scientific Journal of Impact Factor (SJIF): 5.71

e-ISSN (O): 2348-4470
p-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 7, Issue 12, December -2020

Retrieving content based documents by using image processing & NLP through query

Document retrieval system

[1]Prof . Jitendra Musale, [2]Shon Nikam, [3]Akash Deshmukh, [4]Shantanu Gavane, [5]Swapnil Matkar

[1]*Department of Computer Engineering, Ananatrao Pawar College Of Engineering and Research*

A. **Abstract** — *Initially the problem faced was that the documents were searched or retrieved on the basis of the textual annotation which were given to it manually such as meta data that is topic, keyword, date, time which would be easy for computer to understand it and perform retrieval process over annotations. So to overcome such scenario , image processing & NLP can play an important role so that the desired information can be retrieved from the document itself rather using the textual annotations and then with help of NLP we can actually match the textual information. Aim of the system is to retrieve a particular document from database by passing a keyword as a query to it and by comparing the keyword with the actual content of the document it should be retrieved. Objective is to extract particular document from database by comparing its content with keywords which passed as query. To do so we are using image processing and one of the important concept from AI that is Natural Language Processing. Storing paper-based documents converted into digital image format image of effective solution to preserve content in document. Searching the Document Images stored in a repository by using time as a queries in one of importance tasks of document retrieval.*

Keywords- *Optical Character Recognition (OCR), Natural Language Programming (NLP), Image Processing, Artificial Intelligence, Image acquisition, Pre-processing, Pattern Recognition.*

I. INTRODUCTION

Few years back the documents were searched or retrieved on the basis of the textual annotation which were given to it manually such as meta data i.e topic, keyword, date, time which would be easy for computer to understand it and perform retrieval process over annotations. So to overcome such scenario, image processing & NLP can play an important role so that the desired information can be retrieved from the document itself rather using the textual annotations and then with help of NLP we can actually match the textual information. Aim of the system is to retrieve a particular document from database by passing a keyword as a query to it and by comparing the keyword with the actual content of the document it should be retrieved. Objective is to extract particular document from database by comparing its content with keywords which passed as query. To do so we are using image processing and one of the important concept from AI i.e Natural Language Processing. The system aims to find and retrieve content based document from the database by image processing & NLP through query. The system will provide document as per queries provided by user. Attributes which can be assumed to specify the query may be keywords from the actual content of an document .In above statement as it is specified that the keywords which we are focusing on to pass it in a query are not been manually annotated while saving that particular document.

Image processing - Image processing is a method to perform some operations on an image, in order to get an enhanced image or to extract some useful information from it. It is a type of signal processing in which input is an image and output may be image or characteristics/features associated with that image. Most of the offices, organizations, companies are dealing with scanned documents. Document image retrieval is very interesting and efficient way to retrieve information. Process of searching document images with the help of queries is one of the important part in retrieving document. This can be done with help of image processing.

Natural language processing - Natural Language Processing (NLP) is a form of artificial intelligence that helps machine “read” text by simulating the human ability to understand language. One of the important application of NLP is Information Retrieval. As the reader has probably already deduced the complexity associated with natural language is especially key when retrieving textual information to satisfy a user’s information needs. This is why in textual information retrieval, NLP techniques are often use both for facilitating descriptions of document content & for presenting the user’s query, all with the aim of comparing both descriptions and presenting the user documents satisfyingly.

II. LITERATURE WORK

Following are the research papers we studied for the retrieving document using image processing.

1. The Simple Image Processing Scheme for Document Retrieval Using Date of Issue as Query.

- Panuwat Ketwong, Piyabhorn Hongsa-arparsat, Ekkharin Srilaphat, Wilailuck Kaprasit.
- System returning thedesired documents as per the queries provided by users.
- IEEE international conference [2017].

2. Information Processing and Retrieval from CSV File by Natural Language.

- Charmpol Tapsai
- Non-technician users easily retrieve information without the need to learn any additional computer languages or programs.
- IEEE international conference [2019].

3. Image processing technology for text recognition

- Yen-Min Su, Hsing-Wei Peng, Ko-Wei Huang, Chu-Sing Yang.
- This study demonstrates how image-processing technologies can be used in combination with optical character recognition to improve recognition accuracy, efficiency of extracting text from images.
- IEEE international conference [2019].

III. EXISTING SYSTEM

It presents a simple scheme of image processing to retrieve the documents, which contain the desired date of issue printed in Thai alphabets, from a repository. The procedure of this retrieval scheme consists of 4 stages: image acquisition, pre-processing, zone identification, and pattern recognition.In this system their actual focus was to retrieve particular document only on the basis of date of issue rather then any other factor. So the keywords which were passed as a query are dates which are in Thai language and later on they are matched with English language by using template matching technique.

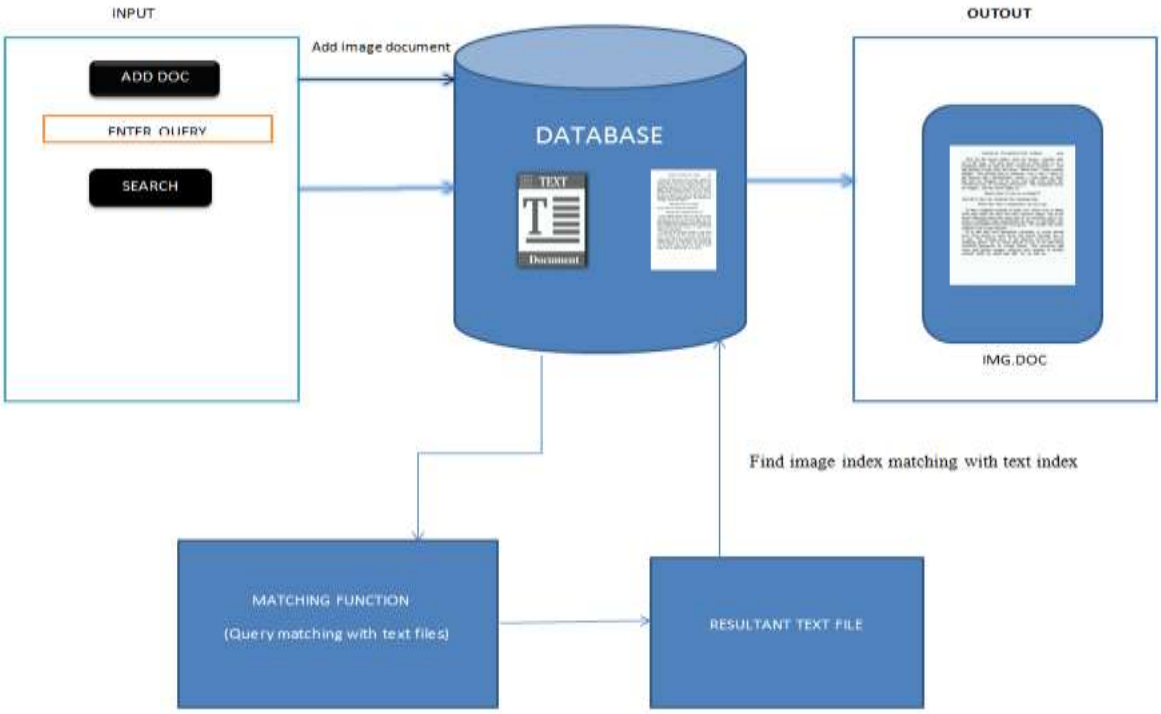
IV. GAPS IN EXISTING SYSTEM

Existing scheme was the retrieval system dealing with query based only on date of issue machine-printed documents.As the documents are been retrieved on the basis of date of issue so there can be multiple documents which are been issued on same date. If there are multiple documents retrieved by using date of issue then we may face one more problem that again we have to find our desired document among documents retrieved.

V. PROPOSED WORK

The gap within the existing system is that it only deals with machine printed documents i.e date printed in it, so in the system which we are proposing we are going to focus on keywords within document content which can be passed in queries to retrieve particular document. Optical Character Recognition or OCR, is a technology that enables you to convert different types of documents, such as scanned paper documents, PDF files or images captured by a digital camera into editable and searchable data.Natural language is especially key when retrieving textual information to satisfy a user's information needs. As information retrieval is an important application of NLP, using this we can compare query with actual content in document. We are developing a document retrieval system with the help of one of image processing technology called OCR. Initially the documents are stored in image format in database.
Previously image documents were retrieved by searching their file name or other textual annotations. In this proposed system we are retrieving these image documents by actual content present in it. Initially when we will store image document in the database, text file will also be created for the same by using OCR and tesseract library. And the keyword which we are going to pass as a query in it will be searched in these text files. If expected keyword is matched with the content of text file then the image document related with that text file will be given as output.

VI. ARCHITECTURE DIAGRAM



MODULES:

Home page – It consists of add button which is used to add image documents to the database. Search bar and search button is used to pass query for searching image document.

Database: It is used to store image documents and text documents. We search document in the database with the help of keyword which is pass as query.

Matching Function: In this module keyword passed by user as query is been searched in text document generated after image processing technique OCR.

Resultant text file: In previous module if the keyword is found in any text document, then that resultant text document is been passed to database to find its matching indexed image document.

Output: If text document received from matching function module finds image document wit same index then that image document is displayed as an output**.**

Technologies :

OCR - When we will store image document in the database, text file will also be created for the same by using **optical character recognition (OCR).** The system takes image as input and OCR converts actual image content to textual format and it is stored in database.

Open CV- Open CV is a cross-platform library using which we can develop real-time computer vision applications. It mainly focuses on image processing, video capture and analysis including features like face detection and object detection. Open CV-Python is a library of Python bindings designed to solve computer vision problems. Open CV-Python makes use of Numpy, which is a highly optimized library for numerical operations.

Numpy- NumPy is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely. NumPy stands for Numerical Python.

Tesseract, Pytesseract – Tesseract is OCR engine. But installing only this is not enough, once we install tesseract we need to connect with python and opencv. Pytesseract library is used to bind tesseract with python, so we can call tesseract from python code.

Python-tesseract is an optical character recognition (OCR) tool for python. That will recognize and "read" the text embedded in images. Python-tesseract is wrapper for **Google tesseract OCR engine**.

VII. APPLICATIONS

- Government offices.
- Banking sector.
- Organisational letter head documents.

VIII. CONCLUSION

In this topic we have seen how the document is been retrieved using image processing and NLP by passing keywords within content of document as query which are machine typed.

REFERENCES

[1] Panuwat Ketwong, Piyabhorn Hongsa-arparsat, Ekkharin Srilaphat, Wilailuck Kaprasit,"The Simple Image Processing Scheme for Document Retrieval Using Date of Issue as Query", IEEE International conference,2017.

[2] Charmpol Tapsai, "Information Processing and Retrieval from CSV File by Natural Language", IEEE International conference,2018.

[3] Yen-Min Su,Hsing-Wei Peng,Ko-Wei Huang,Chu-Sing Yang,"Image processing technology for text recognition", IEEE International conference,2019.

