International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

Content Based Text Retrieval using Optical Character Recognition

Kodamanchili Sirisha
M. Tech, ECE (CESP)
Dr. Y.S.R ANU College
of Engineering & Technology
Acharya Nagarjuna University
Guntur, AP

Dr. G. Prathibha
Asst. Professor, Department of ECE
Dr. Y.S.R ANU College
of Engineering & Technology
Acharya Nagarjuna University
Guntur, AP

***Abstract*— Digital image processing has been increasing exponentially in the last few decades. Its applications range from medicine to entertainment. Almost in every field, digital image processing puts a live effect on things and is growing with time to time and with new technologies. This paper presents a better mechanism to retrieve the text contents present in the images and to obtain more accurate results. Sometimes users are interested in the content present in the image. Content-Based text Retrieval is one of the techniques which extract the required image based on the given text data using the technique Optical Character Recognition and thus it saves the time and minimize the overhead of user to locate the needed information.**

***Keywords*— OCR , metadata.**

I. INTRODUCTION

Although images can be retrieved using metadata such as captions, authors, date, text description, etc., it is not feasible to retrieve the images based on text as it is not easy to define a unique set of keywords to each and every image and when the image database is large it is time consuming to add the keywords. Hence a technical way is introduced to retrieve images based on its content using Optical Character Recognition which is performed on all the images in our database at once and the same is stored in the database which is known as creating meta data and then the third step involves in calculating similarity between the query text and the metadata stored in the database. The result images are the images whose metadata is exactly matched with the query text, and the result images are displayed accordingly.

Optical Character Recognition (OCR) has been a topic of interest for many years. It is defined as the process of digitizing a document image into its constituent characters. Despite decades of intense research, developing OCR with capabilities comparable to that of human still remains an open challenge. Due to this challenging nature, researchers from industry and academic circles have directed their attentions towards Optical Character Recognition. Over the last few years, the number of academic laboratories and companies involved in research on Character Recognition has increased dramatically. The earliest OCR systems were not computers but mechanical devices that were able to recognize characters, but very slow speed and low accuracy. In 1951, M. Sheppard invented a reading robot GISMO that can be considered as the earliest work on modern OCR. GISMO can read musical notations as well as words on a printed page one by one. However, it can only recognize 23 characters. The machine also has the capability to could copy a typewritten page.

II.TYPES OF OCR SYSTEMS

There has been multitude of directions in which research on OCR has been carried out during past years. This section discusses different types of OCR systems have emerged as a result of these researches. These systems can be categorized based on image acquisition mode, character connectivity, font-restrictions etc. Figure 3.5 categorizes the character recognition system. Based on the type of input, the OCR systems can be categorized as

1.Handwriting recognition and
2.Machine printed character recognition.

The former is relatively simpler problem because characters are usually of uniform dimensions, and the positions of characters on the page can be predicted.

A. Handwriting recognition

Handwriting 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. These systems can be divided into two sub-categories i.e. on-line and off-line systems. The former is performed in real-time while the users are writing the character. They are less complex as they can capture the temporal or time based information i.e. speed, velocity, number of strokes made, direction of writing of strokes etc. In addition, there no need for thinning techniques as the trace of the pen is few pixels wide. The offline recognition systems operate on static data i.e. the input is a bitmap. Hence, it is very difficult to perform recognition. There have been many online systems available because they are easier to develop, have good accuracy and can be incorporated for inputs in tablets and PDAs.Maintaining the Integrity of the Specifications.

B. Machine Printed Character Recognition

Machine printed character recognition is the most popular document capturing system at present. It is a process to perform electronic conversion of the text on a physical paper. The text on the document can be either machine print or handwritten.

Machine printed character recognition uses intelligent document recognition technology to read the pattern of the text on the physical paper and convert them accurately in digital format. It increases efficiency and work productivity like never before. It plays the major part in digitization of companies and making workplaces more efficient.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

The earlier OCR system is now trained using Artificial Neural Network through back propagation algorithm.

III. BACK PROPOGATION ALGORITHM

The Back propagation algorithm is a supervised learning method for multilayer feed- forward networks from the field of Artificial Neural Networks. Feed-forward neural networks are inspired by the information processing of one or more neural cells, called a neuron. A neuron accepts input signals via its dendrites, which pass the electrical signal down to the cell body. The axon carries the signal out to synapses, which are the connections of a cell's axon to other cell's dendrites.

The principle of the back propagation approach is to model a given function by modifying internal weightings of input signals to produce an expected output signal. The system is trained using a supervised learning method, where the error between the system's output and a known expected output is presented to the system and used to modify its internal state.

A. STRUCTURE OF PROPOSED SYSTEM

Technically, the back propagation algorithm is a method for training the weights in a multilayer feed-forward neural network. As such, it requires a network structure to be defined of one or more layers where one layer is fully connected to the next layer. A standard network structure is one input layer, one hidden layer, and one output layer.

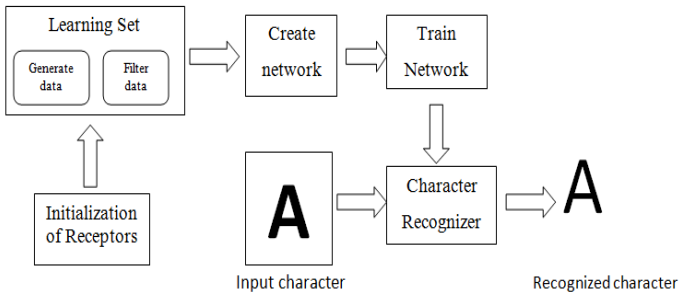


Figure 1 : Structure of Proposed system

The input pattern is presented to the input layer of the network. These inputs are propagated through the network until they reach the output units. This forward pass produces the actual or predicted output pattern, because, back propagation is a supervised learning algorithm, the desired outputs are given as part of the training vector. The actual network outputs are subtracted from the desired outputs and an error signal is produced. This error signal is the basis for the back propagation step, whereby the errors are passed back through the neural network by computing the contribution of each hidden processing unit and deriving the corresponding adjustments needed to produce the correct output. The connection weights are then adjusted and the neural network has just "learned" from an experience. Once the network is trained, it will provide the desired output for any of the input patterns. The network undergoes supervised training, with a finite number of pattern pairs consisting of an input pattern and a desired or target output pattern.

An input pattern is presented at the input layer. The neurons here pass the pattern activations to the next layer neurons, which are in a hidden layer. The outputs of the hidden layer neurons are obtained by the weights and the inputs, these hidden layer outputs become inputs to the output neurons, which process the inputs using an optional bias and a threshold function. The final output of the network is determined by the activations from the output layer.

A similar computation, still based on the error in the output, is made for the connection weights between the input and hidden layers. The procedure is repeated with each pattern pair assigned for training the network. Each pass through all the training patterns is called a cycle or an epoch. The process is then repeated as many cycles as needed until the error is within a prescribed tolerance. The adjustment for the threshold value of a neuron in the output layer is obtained by multiplying the calculated error in the output at the output neuron and the learning rate and the momentum parameter used in the adjustment calculation for weights at this layer.

$$\phi_k = \alpha_s$$

$$Y = f(I) = f\left\{\sum_{i=1}^{n} x_i w_i - \phi_k\right\}$$

After the network has learned the correct classification for a set of inputs from a training set, it can be tested on a second set of inputs to see how well it classifies untrained patterns.

IV. MAJOR PHASES OF OCR

A. Pre-processing Phase:

The aim of pre-processing is to eliminate undesired characteristics or noise in an image without missing any significant information. Preprocessing techniques are needed on color, grey-level or binary document images containing text and/or graphics. Since processing color images is computationally more expensive, most of the applications in character recognition systems utilize binary or grey images. Preprocessing reduces the inconsistent data and noise. It enhances the image and prepares it for the next phases in OCR phases. The effectiveness and easiness for an image can be enhanced to be processed in the next phases by converting the image to the suitable format in the preprocessing phase which is the first phase. Therefore, decreasing the noise that causes the reduction in the character recognition rate is the main important issue in preprocessing phase. Thus, since preprocessing controls the suitability of the input for the successive phases, a primary stage prior to feature extraction phase is the preprocessing phase. Most of the challenges we listed in OCR Challenges' section need to be addressed in preprocessing stage. Some operations that can be considered to carry out can be listed as follows: binarization, noise reduction, skew correction, morphological operations, slant removal, filtering, thresholding, smoothing, compression, and thinning. Some important preprocessing issues with short description were illustrated in Table below.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

Table 1. Some important pre-processing operations

Processes	Description
Binarization	Separates image pixels as text or background.
Noise Reduction	Better improvements of image acquisition devices produced by the advancements in technology.
Skew Correction	Because of the possibility of rotation of the input image through captured image device, document skew should be corrected.
Morphological Operations	Adding or removing pixels to the characters that have holes or surplus pixels.
Thresholding	For an image, separating information from its background.
Thinning and Skeletonisation	Thinning process is the skeletonisation, which regularize the map of the text until reaches most medial one pixel width

B. Segmentation Phase:

The critical and major component of an Optical Character Recognition (OCR) system is the segmentation of text line from images. In general, Text segmentation from a document image merges line segmentation, word segmentation and then character segmentation. Segmentation is the process of isolating text component within an image from the image's background. For appropriate reorganization of the editable text lines from the recognized characters, firstly, segmenting the line of text, then the words are segmented from the segmented line and then from that the characters are segmented. Document segmentation is a major pre-processing phase in implementing an OCR system. It is the process of classifying a document image into homogeneous zones, i.e., that each zone contains only one kind of information, such as text, a figure, a table, or a halftone image. In many cases, the accuracy rate of systems related to the OCR heavily depends on the accuracy of the page segmentation algorithm used. There are three categories of Algorithms of document segmentation [41] As follows:

- Top-down methods,
- Bottom-up methods,
- Hybrid methods.

The top-down approach in a document segments large regions into smaller sub regions recursively. When criterion is met then the document segmentation process will stop and at that stage the ranges obtained constitute the results of final segmentation. But, approaches of bottom-up start by searching for interest pixels and then groups interest pixels. They then manage those interest pixels into connected components that constitute characters which are then combined into words, and lines or text blocks. The integration of both top-down and bottom-up methods is called hybrid approaches. Regarding different aspects of OCR system throughout the last decades many approaches have already been proposed for segmentation.

C.Normalization Phase

As a result of segmentation process isolated characters which are ready to move through feature extraction phase are obtained, hence the isolated characters are minimized to a particular size depending on the algorithms used. The segmentation process is crucial as it converts the image in the form of m*n matrix. These matrices are then commonly normalized by minimizing the size and eliminating the unnecessary information from the image without missing any influential information.

C. Feature Extraction Phase

Feature extraction is the operation of extracting the pertinent features from objects or alphabets to build feature vectors. These feature vectors are then utilized by classifiers to identify the input unit with objective output unit. It becomes effortless for the classifier to classify between dissimilar classes by glancing at these features as it becomes fairly easy to determine. Several techniques are proposed for extracting features from the segmented characters in literature. U. Pal et al have proposed directional chain code features and zoning and for handwritten numeral recognition considered a feature vector of length 100 and have presented a high level of recognition accuracy. But, the feature extraction process is time consuming and complex.

D. Classification Phase

OCR systems broadly utilize the methodologies of pattern recognition, which assigns each example to a predefined class. Classification is the procedure of distributing inputs with respect to detected information to their comparing class in order to create groups with homogeneous qualities, while segregating different inputs into different classes. Classification is conveyed out on the premise of put away features in the feature space, for example, structural features, global features and so forth. It can be said that classification isolates the feature space into several classes taking into account the decision rule. Choosing classifier depends on several agents, such as, number of free parameters, available training set and so forth. Various procedures for OCR are explored by the scientists. Techniques of OCR classification can be categorized as Statistical Techniques, Neural Networks, Template Matching, Support Vector Machine (SVM) algorithms, and Combination of classifier.

E. Post Processing Phase

It has been shown that people can read handwriting by context up to 60%. While preprocessing tries to clean the record in a specific sense, it might evacuate critical data, since the context data is not accessible at this stage. On the off chance that the semantic data were accessible to a specific degree, it would contribute a considerable measure to the precision of the OCR stages. On the other hand, the whole OCR issue is for deciding the context of the saved image. In this way the incorporation of context and shape data in all the phases of OCR frameworks is vital for meaningful upgrades in recognition rates. This is done in the Post processing stage with an input to the early phases of OCR. The least complex method for consolidating the context data is the usage of a dictionary for amending the minor errors of the OCR frameworks. The fundamental thought is to spell check the OCR yield and give a few distinct options for the yields of the recognizer that take place in the dictionary.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021

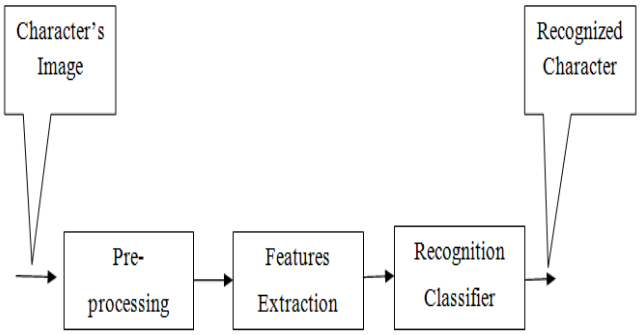


Figure 2. Component of OCR

V. APPLICATIONS OF OCR

A. Invoice Imaging

Invoice imaging is widely used in many business applications to keep track of financial records and prevent a backlog of payments from piling up. In government agencies and independent organizations, OCR simplifies data collection and analysis, among other processes. As the technology continues to develop, more and more applications are found for OCR technology, including increased use of hand writing recognition. Furthermore, other technologies related to OCR, such as barcode recognition used daily in retail and other industries.

B. Banking

Another important application of OCR is in banking, where it is used to process cheques without human involvement. A cheque can be inserted into a machine where the system scans the amount to be issued and the correct amount of money is transferred. This technology has nearly been perfected for printed checks, and is fairly accurate for hand written checks as well reducing the waiting time in banks.

C. Health care

Healthcare has also seen an increase in the use of OCR technology to process paperwork. Healthcare professionals always have to deal with large volumes of forms for patients, including insurance forms as well as general health forms. To keep up with all of this information, it is useful to input relevant data into an electronic database that can be accessed as necessary. Form processing tools, powered by OCR, are able to extract information from forms and put it into databases, so that every patient's data is promptly recorded.

D. Captcha

A CAPTCHA is a program that can generate and grade tests that human can pass but current computer programmers' cannot. Hacking is a serious threat to internet usage, now a day's, most of the human activities like economic transactions, admission for education, registrations, travel bookings etc., are carried out through internet and all this requires a password which is misused by hackers. They create programs to like dictionary attacks and automatic false enrolments which lead to waste of memory and resources of website. Dictionary attack is attack against password authenticated systems where a hacker writes a program to repeatedly try different passwords like from dictionary of most common passwords. In CAPTCHA, an image consisting of series of letters of number is generated which is obscured by image distortion techniques, size and font variation, distracting backgrounds, random segments, highlights and noise in the image. This system can be used to remove this noise and segment the image to make the image tractable for the OCR (Optical Character Recognition) systems.

E. Automatic Number Recognition

Automatic number plate recognition is used as a mass surveillance technique making use of optical character recognition on images to identify vehicle registration plates. ANPR has also been made to store the images captured by the cameras including the numbers captured from license plate. ANPR technology own to plate variation from place to place as it is a region specific technology. They are used by various police forces and as a method of electronic toll collection on pay-per-use roads and cataloguing the movements of traffic or individuals.

F.Handwriting Recognition

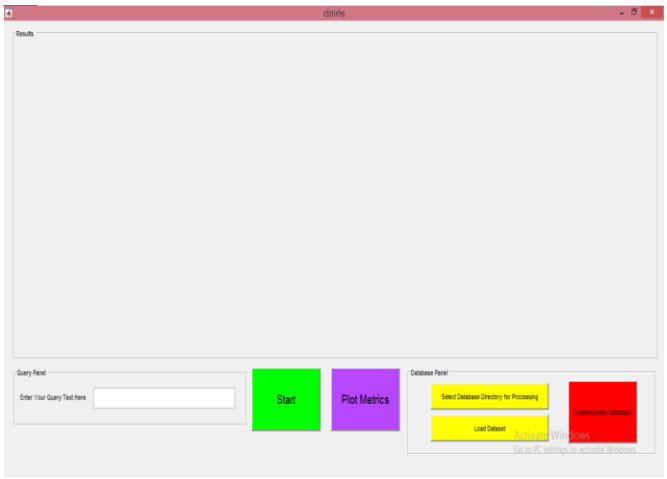
Handwriting recognition is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch-screens and other devices. The image of the written text may be sensed "off line" from a piece of paper by optical scanning (optical character recognition) or intelligent word recognition. Alternatively, the movements of the pen tip maybe sensed "on line", for example by a pen-based computer screen surface.

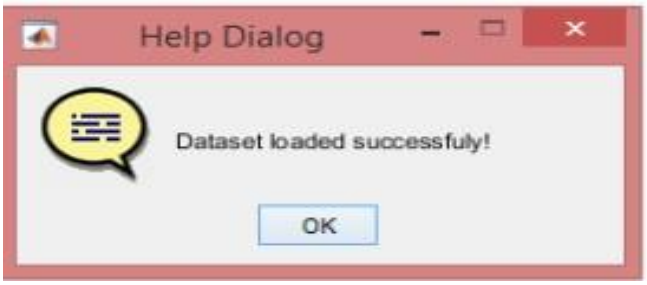
VI. EXPERIMENTAL RESULTS

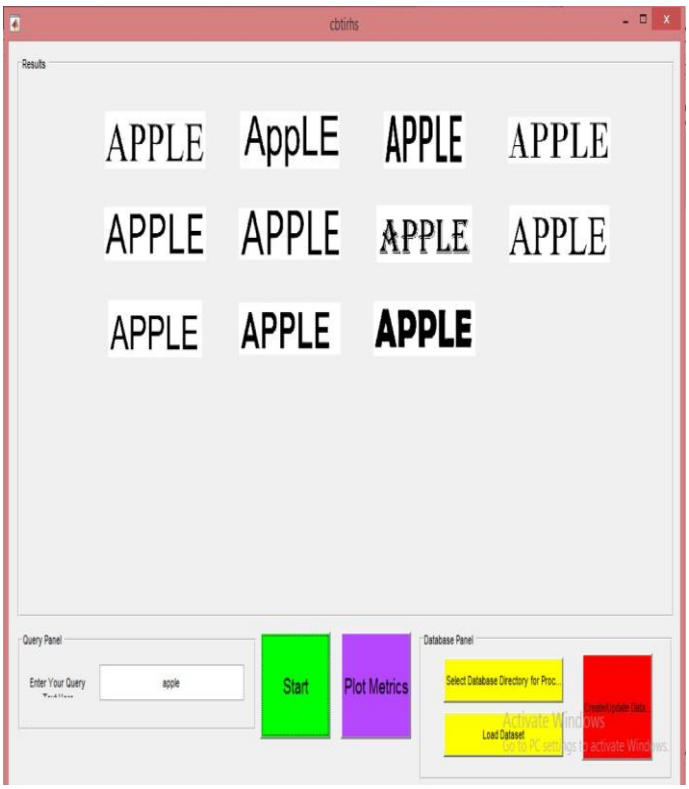
All the functionality is embedded into a figure file. The below user selection screen appears when the figure file or mat file is run. At first we need to create our own database which contains image files, and then process the database file.

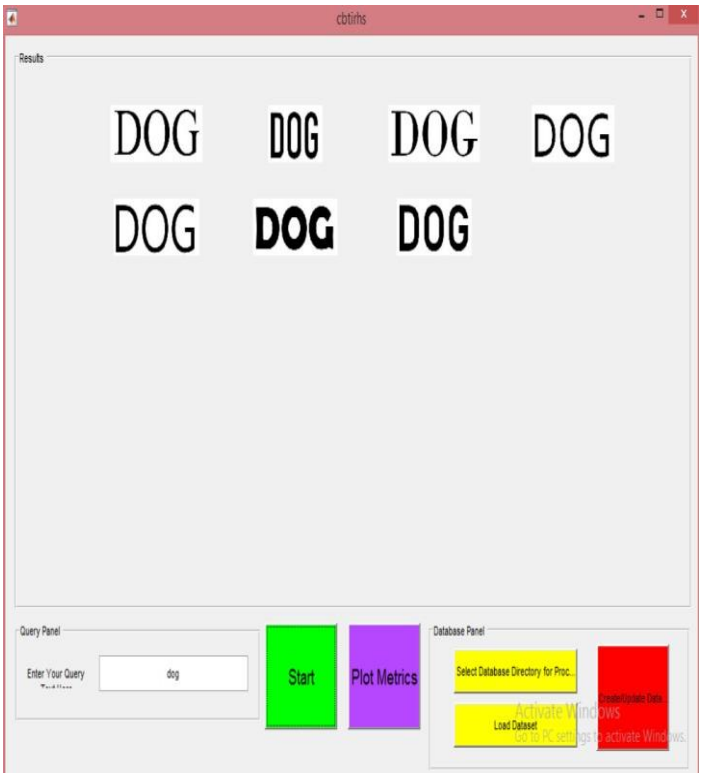
Step 1: Execute the mat file or figure file
Step 2: Click on select database directory for processing
Step 3: Select data base file.
Step 4: click on create or update database.
Step 5: Click on load dataset
Step 6: Select the data base file from my dataset.
Step 7: Give input text in the textbox.ccc
Step 8: Click on start to obtain results.
Step 9:Click plot metrics to obtain precision and recall values.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021









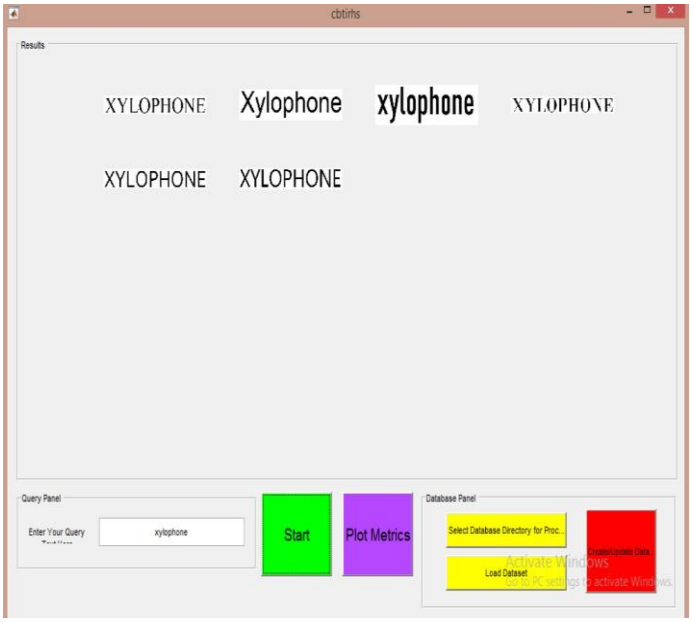
A. Precision and Recall

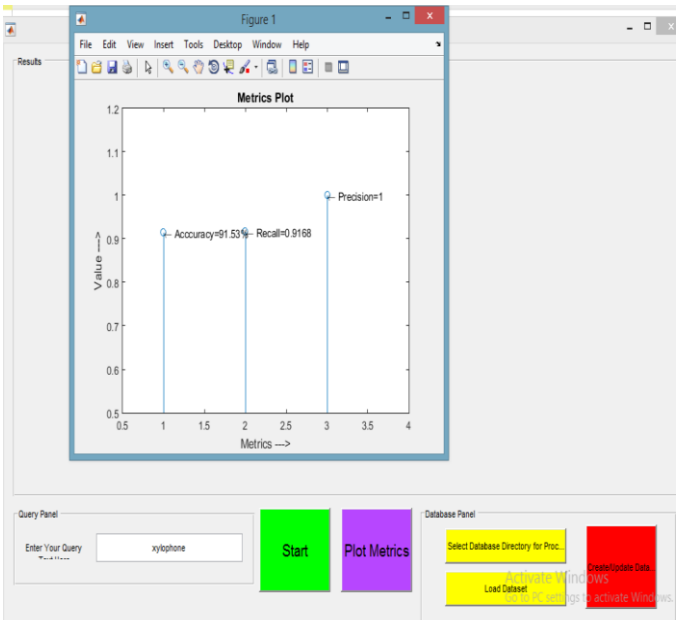
In pattern recognition, information retrieval and classification (machine learning), precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that were retrieved. Both precision and recall are therefore based on relevance.

Precision quantifies the number of positive class predictions that actually belong to the positive class.Recall quantifies the number of positive class predictions made out of all positive examples in the dataset.

If we observe the below metric plot, precision is 1. (the maximum value) since we never get a false output image.This is because of the condition check “display the database image only if it matches with the input query text”. Hence we always get precision 1.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 06, June-2021





VII. CONCLUSION

In this paper, the retrieval of images based on the text has been presented. An OCR is not an atomic process but comprises various phases such as acquisition, pre- processing, segmentation, feature extraction, classification and post-processing. Each of the steps is discussed in detail in this paper. Using a combination of these techniques, further more efficient OCR system can be developed as a future work. The OCR system can also be used in different practical applications such as number-plate recognition, smart libraries and various other real-time applications. Despite of the significant amount of research in OCR, recognition of characters for language such as Arabic, Sindhi and Urdu still remains an open challenge. An overview of OCR techniques for these languages has been planned as a future work. Another important area of research is multi-lingual character recognition system. Finally, the employment of OCR systems in practical applications remains an active area of research.

REFERENCES

[1] Yang, Jufeng, Kai Wang, Jiaofeng Li, Jiao Jiao, and Jing Xu. "A fast adaptive binarization method for complex scene images." In Image Processing (ICIP), 2012 19th IEEE International Conference on, pp. 1889-1892. IEEE, 2012.

[2] Sumetphong, Chaivatna, and SupachaiTangwongsan. "An Optimal Approach towards Recognizing Broken Thai Characters in OCR Systems." Digital Image Computing Techniques and Applications (DICTA), 2012 International Conference on. IEEE, 2012.

[3] AlSalman, AbdulMalik, et al. "A novel approach for Braille images segmentation." Multimedia Computing and Systems (ICMCS), 2012 International Conference on. IEEE, 2012.

[4] Mutholib, Abdul, Teddy Surya Gunawan, and Mira Kartiwi. "Design and implementation of automatic number plate recognition on android platform." Computer and Communication Engineering (ICCCE), 2012 International Conference on. IEEE, 2012.

[5] Chi, Bingyu, and Youbin Chen. "Reduction of Bleed-through Effect in Images of Chinese Bank Items." Frontiers in Handwriting Recognition (ICFHR), 2012 International Conference on. IEEE, 2012.

[6] Ramakrishnan, Kandan, and Evgeniy Bart. "Learning domain-specific feature descriptors for document images." Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on. IEEE, 2012.

[7] Chattopadhyay, T., Ruchika Jain, and Bidyut B. Chaudhuri. "A novel low complexity TV video OCR system." Pattern Recognition (ICPR), 2012 21st International Conference on. IEEE, 2012.