PROJECT REPORT ON

OCR Data Extraction

SUBMITTED BY

Sagar Jamunaprasad Gupta

MSc. Information Technology

ACADEMIC YEAR 2020-2021

Abstract

This Digitalization of documents is now being done in all fields to reduce paper usage. The availability of modern technology in the form of scanners and cameras supports the growth of multimedia data, especially documents stored in the form of image files. Searching a particular text in a large-scale scanned document images is a difficult task if the document is in the form of images where the text has not been extracted. In this research, text extraction method of large-scale scanned document images using Google Vision OCR is proposed. The object of research is student thesis documents, which includes the cover page, the approval page, and abstract.All documents are stored in the university's digital library. Extraction process begins with preparing the input folder that contains image documents (in JPEG format). The image document is then extracted using Google Vision OCR in order to obtain text document (in TXT format) The same process is repeated for the entire documents in the folder.Test results have shown that the proposed methods were able to extract all test documents successfully. Google Vision OCR also shows better extraction performance compared to other OCR tools. The proposed automated extraction systems can recognize text in alarge-scale image document accurately and can be operated in a real-time environment. In many alternative fields, there's a high demand for storing information to a memory device disk from the info available in printed or handwritten documents or images to later re-utilize this information by means of computers. One simple thanks to store information to a system from these printed documents might be first to scan the documents so store them as image files. But to re-utilize this information, it'd very difficult to read or query text or other information from these image files. Therefore a technique to automatically retrieve and store information, particularly text, from image files is required. Optical character recognition is an active research area that attempts to develop a ADPS with the flexibility to extract and process text from images automatically. The objective of OCR is to realize modification or conversion of any sort of text or text-containing documents like handwritten text, printed or scanned text images, into an editable digital format for deeper and further processing. Therefore, OCR enables a machine to automatically recognize text in such documents. Some major challenges have to be recognized and handled so as to attain a successful

automation. The font characteristics of the characters in paper documents and quality of images are just some of the recent challenges. because of these challenges, characters sometimes might not be recognized correctly by system. during this paper we investigate OCR in four alternative ways. First we provides a detailed overview of the challenges that may emerge in OCR stages. Second, we review the overall phases of an OCR system like pre-processing, segmentation, normalization, feature extraction, classification and post-processing. Then, we highlight developments and main applications and uses of OCR and eventually, a quick OCR history are discussed. Therefore, this discussion provides a awfully comprehensive review of the state-of-the-art of the sector.

ACKNOWLEDGEMENT

I'm so glad to present my project report on **"OCR Data Extraction",** which is prepared as a part of my Third semester project in "Masters of Science in Information Technology “This project work is the outcome of guidance provided by dedicated and experienced personnel of "S. K. SOMAIYA DEGREE COLLEGE OF ARTS, SCIENCE AND COMMERCE". We are thankful to God and our parents for helping us to get a good education.

With profound sense of indebtedness, we use this opportunity to express our sincere thanks to our beloved principal “Dr. Manali Londhe", for providing all support and necessary facilities throughout the course and for the inspiration to complete my project. We express our sincere gratitude to “Dr. Rajendra Patil" Head, “Marielia Assumption” coordinator of Masters of Science in Information Technology Department for inspiration and permission to carry out our project. We wish to extend our thanks to our project guide “Dr. Rajendra Patil” and Subject guides "Dr. Hiren Dand", “Dr.Abhijeet Kale”.

Last but not the least we extend our sincere thanks to all our classmates for their help and support. And also we would like to extend our deep sense of gratitude to all those who have assisted us in the successful completion of this project.

Sagar Jamunaprsad Gupta

Thanking You.

List of Tables

Introduction

It is natural and accustomed that we should demand to build and design machines that can recognize patterns. From automated optical character recognition to face recognition, fingerprint identification, speech recognition, DNA sequence identification and much more, it is clear that accurate and reliable pattern recognition by machine would be greatly useful. Optical character recognition is an active research area that attempts to develop a computer system with the ability to extract and process text from images automatically. These days there is a huge demand for storing information to a computer storage disk from the data available in printed or handwritten documents to later re-utilize this information by means of computers. One simple way to store information to computer system from these paper documents could be to first scan the documents and then store them as image files. But to re-utilize this information, it would very difficult to read or query text or other information from these image files. Therefore a technique to automatically retrieve and store information, in particular text, from image files is needed. Of course, this is not a very trivial task. Some major challenges need to be laid out and handled in order to achieve a successful automation. The font characteristics of the characters in paper documents and quality of images are only some of the recent challenges. Due to these challenges, characters sometimes may not be recognized correctly by computer system. Thus there is a need of mechanisms of character recognition to perform Document Image Analysis (DIA) which overcomes these challenges and produces electronic format from the transformed documents in paper format [2]. Similarly, Optical Character Recognition (OCR) is the process of modification or conversion of any form of text or text-containing documents such as handwritten text, printed or scanned text images, into an editable digital format for deeper and further processing. Optical character recognition technology enables a machine to automatically recognize text in such documents. In real world example, it is like combination of mind and eye of human body. An eye can detect, view and extract the text from the images but absolutely the human's brain processes that detected or extracted text read by eye [1]. Of course OCR technology has not advanced enough to compete with human's ability. The performance and accuracy of OCR is directly dependent upon the quality of input documents.

Again, when we think of human’s ability to recognize text, the performance of brain’s process directly depends upon the quality of the input read by eye. While designing and implementing a computerized OCR system, several problems and challenges can occur. For example there is very slight difference between some digits and letters for computers to recognize them and distinguish one from the others correctly. For example, it may not be very easy for computers to differentiate between digit “0” and letter “o”, especially when these characters are embedded in a very dark and noisy background. One of the main focuses of OCR research has been to recognize cursive scripts and handwritten text for its broad application area. Today, to solve the text recognition problem several different types of OCR software exist such as Desktop OCR, Server OCR, web OCR and so on. Since the OCR research is an active and important field in general pattern recognition problems, due to its fast progress, comprehensive reviews of the field are needed on a regular basis to keep track of the new advancements. One such review was published to discuss the challenges with text recognition in scene imagery [2]. This paper attempts to elaborate on these kinds of studies by providing a comprehensive literature review of optical character recognition research. We discuss major challenges and main phases of optical character recognition such us preprocessing, segmentation, normalization, feature extraction, classification and post processing in detail which needs to be considered during implementing any application related to the OCR, and in the last section of our paper some OCR application. This Digitalization of documents is now being done in all fields to reduce paper usage. The availability of modern technology in the form of scanners and cameras supports the growth of multimedia data, especially documents stored in the form of image files. Searching a particular text in a large-scale scanned document images is a difficult task if the document is in the form of images where the text has not been extracted. In this research, text extraction method of large-scale scanned document images using Google Vision OCR is proposed. The object of research is student thesis documents, which includes the cover page, the approval page, and abstract. All documents are stored in the university's digital library. Extraction process begins with preparing the input folder that contains image documents (in JPEG format). The image document is then extracted using Google Vision OCR in order to obtain text document (in TXT format) The same process is repeated for the entire documents in the folder. Test results have shown that the proposed methods were able to extract all test documents successfully. Google Vision OCR also shows better extraction performance compared to other OCR tools. The proposed automated extraction systems can recognize text in a large-scale image document accurately and can be operated in a real-time environment.

What is OCR on a scanner?

Optical Character Recognition (OCR) is a technology that allows you to extract data from scanned documents resulting in a text which you can then edit, update, or aggregate with other tools for data analysis and a range of other uses.

Optical Character Recognition (OCR), is essentially the conversion of scanned images with text, be it typed, in print, or written by hand, into … well … text. Typically you see OCR used in extracting text information from photos, passports, and scanned documents. OCR is often used for "digitizing" recognized text, so it can be utilized later, edited, searched, aggregated for analysis, etc.

How does OCR Data Extraction works?

A scanner merely takes a picture of a document. Whatever type of paper document you started with, it becomes an image consisting of dots and lines – or unstructured data – that an ECM cannot read. Without OCR, the scanned document can be stored, retrieved and reviewed, but the data is unusable without OCR data extraction.

With Optical Character Recognition software, the unstructured data is converted to structured, usable data the moment a document is scanned or received electronically. The OCR software identifies and extracts letters from the image and assembles them into words and sentences, essentially translating those dots and lines that the ECM couldn't read into "structured" data in the form of a readable, editable document. These documents include Word, PDF, Excel and other text formats.

Since some industrial scanners can scan up to 120 pages per minute, an ECM with OCR software can process data at a rate significantly faster than a human employee.

What is optical character recognition (OCR) and why is it important for RPA?

Optical character recognition (OCR) is a key feature of any good robotic process automation (RPA) solution. In short, OCR is a technology used to extract text from images and documents via mechanical or electronic means. It converts typed, handwritten or printed text into machine-encoded text – this data can then be used in electronic business processes without someone manually capturing it.

OCR has been around in various forms for more than 100 years, but unlike the earlier versions of the technology that need to be trained one font at a time with images of each character, today’s artificial intelligence (AI) powered OCR solutions can recognize and capture data from machine printed documents with high levels of accuracy. Their ability to accurately decipher handwritten text is also rapidly improving

Data Extraction

It is that the act or method of retrieving information out of (usually unstructured or poorly structured) information sources for additional processing or information storage (data migration). The import into the intermediate extracting system is so typically followed by information transformation and probably the addition of information before export to a different stage within the information progress.

Usually, the term information extraction is applied once (experimental) information is initial foreign into a laptop from primary sources, like measurement or recording devices. Today's electronic devices can typically gift associate electrical connecter (e.g. USB) through that 'raw data' is streamed into a private laptop.

Data sources

Typical unstructured information sources embrace web content, emails, documents, PDFs, scanned text, mainframe reports, spool files, classifieds, etc. that is additional used for sales or selling leads. Extracting information from these unstructured sources has grownup into a substantial technical challenge wherever as traditionally information extraction has had to alter changes in physical hardware formats, the bulk of current information extraction deals with extracting information from these unstructured information sources, and from totally different code formats. This growing method {of information|of knowledge of information extraction[1] from the net is stated as "Web data extraction" or "Web scraping".

Imposing structure

The act of adding structure to unstructured information takes variety of forms

• Using text pattern matching like regular expressions to spot little or large-scale structure e.g. records in a very report and their associated information from headers and footers;

• Using a table-based approach to spot common sections inside a restricted domain e.g. in emailed resumes, characteristic skills, previous work expertise, qualifications etc. employing a commonplace set of ordinarily used headings (these would dissent from language to language), e.g. Education may be found underneath Education/Qualification/Courses;

• Using text analytics to aim to grasp the text and link it to alternative info

Types of Data extraction

Extraction jobs is also scheduled , or analysts could extract knowledge on demand as settled by business desires and analysis goals. knowledge are often extracted in 3 primary ways:

Update notification

The easiest thanks to extract knowledge from a supply system is to own that system issue a notification once a record has been modified. Most infos offer a mechanism for this so they will support knowledgebase replication (change data capture or binary logs), and plenty of SaaS applications offer webhooks, which provide conceptually similar practicality.

Incremental extraction

Some knowledge sources ar unable to produce notification that associate degree update has occurred, however they're ready to establish that records are changed and supply associate degree extract of these records. throughout ensuant ETL steps, the info extraction code must establish and propagate changes. One disadvantage of progressive extraction is that it should not be ready to observe deleted records in supply knowledge, as a result of there's no thanks to see a record that's not there.

Full extraction

The first time you replicate any supply you have got to try to to a full extraction, and a few knowledge sources don't have any thanks to establish knowledge that has been modified, thus reloading a full table is also the sole thanks to get knowledge from that supply. as a result of full extraction involves high knowledge transfer volumes, which might place a load on the network, it's not the simplest possibility if you'll avoid it.

The data extraction method

Whether the supply may be a info or a SaaS platform, the info extraction method involves the subsequent steps:

1. Check for changes to the structure of the info, together with the addition of latest tables and columns. modified knowledge structures need to be controlled programmatically.

2. Retrieve the target tables and fields from the records such by the integration's replication theme.

3. Extract the suitable knowledge, if any.

Extracted knowledge is loaded into a destination that is a platform for metallic element reportage, like a cloud knowledge warehouse like Amazon Redshift, Microsoft Azure SQL knowledge Warehouse, Snowflake, or Google BigQuery. The load method must be specific to the destination.

API-specific challenges

While it should be attainable to extract knowledge from a info victimization SQL, the extraction method for SaaS merchandise depends on every platform's application programming interface (API). operating with arthropod genus are often challenging:

• APIs ar totally different for each application.

• Many arthropod genus don't seem to be well documented. Even arthropod genus from respectable, developer-friendly firms typically have poor documentation.

- APIs amendment over time. as an example, Facebook’s “move quick and break things” approach means that the corporate ofttimes updates its reportage arthropod genus – and Facebook doesn’t continually send word API users prior to.

Data Extraction Techniques

How to extract data

News stories of information being ‘amassed’ or ‘mined’ area unit common. But, wherever is all the info coming back from, and the way is it being processed for valuable use? to know the info trade, analyzing the info extraction method is important. helpful capital information must be within the variety of client identities, shopper behaviors, beliefs, and different customer-related data. individuals area unit at the core of information. firms making an attempt to travel ‘smart’ have to be compelled to generate and extract reams of valuable information.

Extraction strategies in information Warehouses

Extraction collects information from supply systems. The extracted information is unbroken in a very ‘data warehouse’ for any probe. The tools utilized in this transfer and transformation of information area unit - Extract, Transform, and cargo (ETL). Extraction is that the initiative. Here’s however Extraction strategies operate in information warehouse environments --

· the information the info the information} is reworked and keep within the data warehouse.

· information warehouses use group action process applications as supply systems. for instance, a group action process application might contain sales analytics information. This warehouse becomes the supply system for the company’s information analyst.

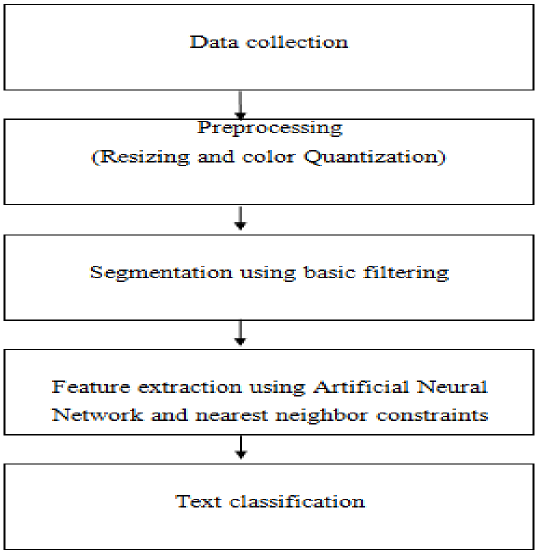
· Extraction is that the most intricate task within the ETL. Most supply systems area unit inadequately recorded. determinant the worth of information (in terms of its eligibility for extraction) may be a advanced method.

· information extraction may be a continuous method. The warehouses have to be compelled to be updated as per the incorporation of latest information within the supply systems.

1.1. Objective:-

The objective of OCR is to realize modification or conversion of any sort of text or text-containing documents like handwritten text, printed or scanned text images, into an editable digital format for deeper and further processing. Therefore, OCR enables a machine to automatically recognize text in such documents. Some major challenges have to be recognized and handled so as to attain a successful automation. The font characteristics of the characters in paper documents and quality of images are just some of the recent challenges.

A database of characters will be formed by collecting a wide range of scanned printed alphabets(A-Z).Input image will be de noised and enhanced using pre-processing step. Image processing using basic filtering.


Data collection
Preprocessing
(Resizing and color Quantization)
Segmentation using basic filtering
Feature extraction using Artificial Neural Network and nearest neighbor constraints
Text classification

1.1.1 Purpose

If an organization needs to change any document into editable digital layout then there is not anything better than OCR services. Optical Character Recognition process saves the time and attempt of raising a digital model of any document. Alterations are taking place in the field of Data Doorway service with the improvement of modern technology. No need to type characters manually on a digital categorizer. OCR application can not only read the fonts but also discern line breaks in a document. The procedure of OCR is easy to manage. The entire process of converting hard copy of a document into electronic document may take only a few seconds. (R2) Just place the hard copy of the document inside a scanner and get the digital format of that document with the help of OCR.

Advanced Features

The advanced features of the online booking system to be implemented were as follows:

1. Take into thought Human Computer Interaction (HCI) issues relating to the ease for users to actually make bookings. This would mainly involve painstaking analysis into the design of the graphical user interface (GUI).
2. Suitable security issue needed to be recognized and thus determined

1.1.2. **Scope and Applicability:**

A number of techniques that are used for optical character recognition have been discussed which uses correlation and neural networks. Much other advancement in Optical Character Recognition are being under development. The paper presents a brief survey of the applications in various fields along with experimentation into few selected fields. The proposed method is extremely efficient to extract all kinds of bimodal images including blur and illumination. The paper will act as a good literature survey for researchers starting to work in the field of optical character recognition. The reason of its complexities are its characters shapes, its top bars and end bars more over it has some modified, vowel and compound characters and also one of the important reasons for poor recognition in OCR system is the error in character recognition

2. Review of Literature

Literature appraisal is an evaluative statement of information found in the prose related to the selected area of study. The appraisal should describe summing up, assess and clarify the selected theme prose. It gives a hypothetical pedestal for the research and helps (the author) determine the natural world of your research. Works which are immaterial should be surplus and those which are tangential should be looked at critically.

A prose review is more than the search for information, and goes further than being an evocative annotated bibliography. All works built-in in the review must be read, evaluated and analyzed (which you would do for an gloss bibliography), but relationships between the prose must also be recognized and spoken, in relation to your field of research. R (9)

"In writing the prose review, the purpose is to put across to the reader what knowledge and ideas have been established on a topic, and what their potency and weaknesses are. The prose review must be defined by a guiding idea (e.g. your research objective, the difficulty or issue you are discussing or your argumentative notion). It is not just a evocative list of the material available, or a set of summaries‟ (R 10)

Optical character recognition (OCR) is an significant research area in outline recognition. The purpose of an OCR system is to be familiar with alphabetic letters, numbers, or other characters, which are in the outward appearance of digital images, without any human interference [R 11]. This is accomplished by searching a match between the skin tone. Ideally, we would like the features to be separate for diverse character descriptions so that the computer can take out the right model from the library without some confusion. At the same time, we also want the skin tone to be robust enough so that they will not be exaggerated by viewing transformations, noises, resolution variations and other issues.The below figure exemplify the basic processes of an OCR scheme.

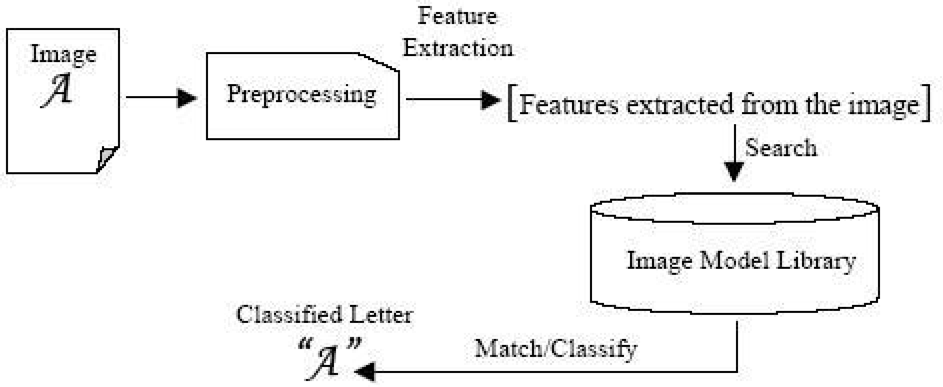


Figure 3: How a character is recognized by OCR

This technology allows a machine to routinely recognize characters through an optical apparatus. Human beings be familiar with many things in this manner our eyes are the "optical mechanism." But while the brain "sees" the input, the ability to understand these signals varies in every person according to many factors. By reviewing these variables, we can appreciate the challenges faced by the technologist mounting an OCR system.

First, if we understand writing a page in a language other than our own, we may be familiar with the various characters, but be unable to be familiar with words. However, on the same page, we are usually able to understand arithmetical statements - the signs for figures are universally used. This explains why many OCR systems recognize figures only, while relatively few appreciate the full alphanumeric character range.

Second, there is resemblance between many numerical and alphabetical symbol forms. For example, while investigative a string of characters combining letters and figures, there is very little visible difference between a capital letter "O" and the numeral "0." As humans, we can re-understand writing the sentence or whole section to help us determine the accurate meaning. This procedure, however, is much more difficult for a machine. Third, we rely on dissimilarity to help us be familiar with characters. We may find it very hard to read text which come into views against a very dark background, or is written over other words or graphics. Again, programming a system to appreciate only the relevant data and ignore the rest is a difficult task for OCR engineers.

There are many other troubles which confront the developers of OCR systems. In this paper, we will review the history, progressions, aptitude and limitations of existing schemes. This analysis should help to decide if OCR is the right application for your company's needs, and if so, which type of system to put into practice.

Image Acquisition:

Digitized/Digital Image is initially taken as input. The most common of these devices is the electronic tablet or digitizer. These devices use a pen that is digital in nature. Input images for handwritten characters can also be taken by using other methods such as scanners, photographs or by directly writing in the computer by using a stylus.

A. Preprocessing:

Pre-processing is the basic phase of character recognition and it's crucial for good recognition rate. The main objective of pre-processing steps is to normalize strokes and remove variations that would otherwise complicate recognition and reduce the recognition rate. These variations or distortions include the irregular size of text, missing points during pen movement collections, jitter present in text, left or right bend in handwriting and uneven distances of points from neighbouring positions. Pre-processing includes five common steps, namely, size normalization and centering, interpolating missing points, smoothing, slant correction and resampling of points.

B. Segmentation:

Segmentation is done by separation of the individual characters of an image. Generally document is processed in a hierarchical way. At first level lines are segmented using row histogram. From each row, words are extracted using column histogram and finally characters are extracted from words.

C. Feature Extraction:

The main aim of feature extraction phase is to extract that pattern which is most pertinent for classification. Feature extraction techniques like Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), Chain Code (CC), Scale Invariant Feature Extraction (SIFT), zoning, Gradient based features, Histogram might be applied to extract the features of individual characters. These features are used to train the system.

D. Classification:

When input image is presented to HCR system, its features are extracted and given as an input to the trained classifier like artificial neural network or support vector machine. Classifiers compare the input feature with stored pattern and find out the best matching class for input.

E. Post Processing:

Post-processing refers to the procedure of correcting misclassified results by applying linguistic knowledge. Post processing is processing of the output from shape recognition. Language information can increase the accuracy obtained by pure shape recognition. For handwriting input, some shape recognizers yield a single string of characters, while others yield a number of alternatives for each character, often with a measure of confidence for each alternative.

2.1. Related work

Claudiu et al. (2011) [1] has investigated using simple training data preprocessing gave us experts with errors less correlated than those of different nets trained on the same or bootstrapped data. Hence committees that simply average the expert outputs considerably improve recognition rates. Our committee-based classifiers of isolated handwritten characters are the first on par with human performance and can be used as basic building blocks of any OCR system (all our results were achieved by software running on powerful yet cheap gaming cards).

Georgios et al. (2010) [2] has presented a methodology for off-line handwritten character recognition. The proposed methodology relies on a new feature extraction technique based on recursive subdivisions of the character image so that the resulting sub-images at each iteration have balanced (approximately equal) numbers of foreground pixels, as far as this is possible. Feature extraction is followed by a two-stage classification scheme based on the level of Vol. 4, No. 6 June 2013 ISSN 2079-8407 Journal of Emerging Trends in Computing and Information Sciences ©2009-2013 CIS Journal. All rights reserved. http://www.cisjournal.org 547 granularity of the feature extraction method. Classes with high values in the confusion matrix are merged at a certain level and for each group of merged classes, granularity features from the level that best distinguishes them are employed. Two handwritten character databases (CEDAR and CIL) as well as two handwritten digit databases (MNIST and CEDAR) were used in order to demonstrate the effectiveness of the proposed technique.

Sankaran et al. (2012) [3] has presented present a novel recognition approach that results in a 15% decrease in word error rate on heavily degraded Indian language document images. OCRs have considerably good performance on good quality documents, but fail easily in presence of degradations. Also, classical OCR approaches perform poorly over complex scripts such as those for Indian languages. Sankaran et al. (2012) [3] addressed these issues by proposing to recognize character n-gram images, which are basically groupings of consecutive character/component segments. Their approach was unique, since they use the character n- grams as a primitive for recognition rather than for post- processing.

By exploiting the additional context present in the character n-gram images, we enable better disambiguation S between confusing characters in the recognition phase. The labels obtained from recognizing the constituent n-grams are then fused to obtain a label for the word that emitted them. Their method is inherently robust to degradations such as cuts and merges which are common in digital libraries of scanned documents. We also present a reliable and scalable scheme for recognizing character n-gram images. Tests on English and Malayalam document images show considerable improvement in recognition in the case of heavily degraded documents.

Jawahar et al. (2012) [4] has propose a recognition scheme for the Indian script of Devanagari. Recognition accuracy of Devanagari script is not yet comparable to its Roman counterparts. This is mainly due to the complexity of the script, writing style etc. Our solution uses a Recurrent Neural Network known as Bidirectional Long- Short Term Memory (BLSTM). Our approach does not require word to character segmentation, which is one of the most common reason for high word error rate. Jawahar et al. (2012) [4] has reported a reduction of more than 20% in word error rate and over 9% reduction in character error rate while comparing with the best available OCR system.

K. Gaurav, Bhatia P. K. [5] Et al, this paper deals with the various pre-processing techniques involved in the character recognition with different kind of images ranges from a simple handwritten form based documents and documents containing colored and complex background and varied intensities. In this, different preprocessing techniques like skew detection and correction, image enhancement techniques of contrast stretching, binarization, noise removal techniques, normalization and segmentation, morphological processing techniques are discussed. It was concluded that using a single technique for preprocessing, we can’t completely process the image. However, even after applying all the said techniques might not possible to achieve the full accuracy in a preprocessing system.

Salvador España-Boquera et al [6], in this paper hybrid Hidden Markov Model (HMM) model is proposed for recognizing unconstrained offline handwritten texts. In this, the structural part of the optical model has been modelled with Markov chains, and a Multilayer Perceptron is used to estimate the emission probabilities.

In this paper, different techniques are applied to remove slope and slant from handwritten text and to normalize the size of text images with supervised learning methods. The key features of this recognition system were to develop a system having high accuracy in preprocessing and recognition, which are both based on ANNs.

In [7], a modified quadratic classifier based scheme to recognize the offline handwritten numerals of six popular Indian scripts is proposed. Multilayer perceptron has been used for recognizing Handwritten English characters

[8]. The features are extracted from Boundary tracing and their Fourier Descriptors. The character is identified by analysing its shape and comparing its features that distinguish each character. Also an analysis has been carried out to determine the number of hidden layer nodes to achieve high performance of the back propagation network.

Types of information Extraction Tools

Data engineers whereas coming up with this advanced method have very important choices concerning

1. The tactic of extraction

2. A way to clean and remodel the info for any processing?

In terms of Extraction strategies, there area unit 2 choices – Logical and Physical.

Logical Extraction additionally has 2 choices - Full Extraction and progressive Extraction.

I. Logical

Full Extraction

All information is extracted directly from the supply system quickly. There's no would like for extra logical/technological data (for instance, dates of once the supply system was updated). as an example, to export one file concerning a worth amendment, the system fully extracts the organization's money records (copying the whole supply table).

Incremental Extraction

Incremental Extraction deals with delta changes within the information. The Extraction tool is conscious of its have to be compelled to acknowledge new or modified data supported time and dates. victimization the progressive Extraction methodology implies that the info engineer can need to add advanced extraction logic 1st to the supply systems.

II. Physical Extraction

Source systems typically have bound restrictions or limitations. as an example, drawing information from out-of-date information storage systems via logical extraction is not possible. the info will be extracted solely by Physical Extractions. the 2 kinds of physical extraction embody - on-line and Offline Extraction.

Online Extraction

The online information extraction method involves direct information transference from the supply system to the info warehouse. For this method to be practical, the extraction tools have to be compelled to directly connect either to the supply system or shift system that options pre-configured information chambers. The shift system is a precise copy of the supply system, except that the info is additional structured.

Offline Extraction

There's no direct extraction from the supply system. the method takes place outside the supply system. the info in such processes is either already structured (for instance, entry/exit logs) or structured via extraction routines. The scope of information that must be extracted {and the|and therefore the|and additionally the} introduce that the ETL method is working at that point also influences the determination of a way to extract. primarily, businesses can need to invest in each logical and physical information extraction strategies.

Data Capture

Data capture is a complicated extraction method. It permits the extraction of information from documents, changing it into machine-readable information. This method is employed to gather vital structure data once the supply systems area unit within the variety of paper/electronic documents (receipts, emails, contracts, etc.).

New information Capture Systems incorporate the employment of optical character recognition tools. information that's scanned from digital documents area unit regenerate into machine-readable information (and sent to the info warehouse for any processing). machine-driven information capture processes play a vital role in desegregation ancient businesses into the fold. These systems scale back the necessity for tedious labor, like manual information entry. The processes area unit quicker and additional efficient within the long-term. With the assistance of information Capture, businesses will currently apace transfer their structure content into sensible processes. fashionable information capture tools will currently even produce logical maps so users will visually select their extraction approach. You get to be told additional regarding the operations in your information warehouse due to the easy interface of information capture tools.

COMPARISION BETWEEN DIFFERENT TECHNIQUES

Sr.no	Method	Accuracy	purpose
1.	Hand printed symbol recognition.	97% overall.	Extract the geometrical, topological and local measurements required to identify the character.

2.	OCR for cursive handwriting	88.8% for lexicon size 40,000.	the character. OCR for cursive handwriting. [6] 88.8% for lexicon size 40,000. To implement segmentation and recognition algorithms for cursive handwriting
3	Recognition of handwritten numerals based upon fuzzy model.	95% for Hindi and 98.4% for English numerals overall.	The aim is to utilize the fuzzy technique to recognize handwritten numerals for Hindi and English numerals
4.	Combining decision of multiple connectionist classifiers for Devanagari numeral recognition.	89.6% overall.	To use a reliable and an efficient technique for classifying numerals.
5.	Hill climbing algorithm for handwritten character recognition.	93% for uppercase letters.	To implement hill climbing algorithm for selecting feature subset.

6.	Optimization of feature selection for recognition of Arabic characters	88% for numbers and 70% for letters.	To apply a method of selecting the features in an optimized way.
7.	Handwritten numeral recognition for six popular Indian scripts.	99.56% for Devanagari, 98.99% for Bangla, 99.37% for Telugu, 98.40% for Oriya, 98.71% for Kannada and 98.51% for Tamil overall.	To find out the recognition rate for the six popular Indian scripts.

PROPOSED SYSTEM

We proposed a system here for recognition of characters from images with the help of Artificial Neural Network and nearest neighbour approach. The main parts of the consideration are the data collection, pre processing step, then the extraction of symbol of interest from the input scanned image, after that conversion of this identified or located symbol into a pixel matrix of specific or standard size. Then the next part is the classification of the symbol under consideration which can be done by assigning the different loads to the layer of neural network. In case of data collection, the hidden layer neurons have to assign weights for the classification of characters according the specific font style. Nearest Neighbour Approach: It is used for classification purposes .Here, for every component, two consecutive nearest neighbours are found by the application of Euclidean distance. The different parameters like dimensions, distance and the alignment of component and their neighbours are compared.

3. Software Requirement and Specification

A software requirements specification (SRS) is a description of a software system to be developed. It is modeled after business requirements specification (CONOPS), also known as a stakeholder requirements specification (STRS). The software requirements specification lays out functional and non-functional requirements, and it may include a set of use cases that describe user interactions that the software must provide to the user for perfect interaction.

3.1 Software and Hardware Requirement

3.1.1 Software Requirement

- Operating System: Windows 7 or later versions
- Software: UiPath Studios (Community Edition in this case)
  Web Browser (For Web Scrapping)
  Microsoft Office (For Data Import and Export)
- Plugins: Microsoft Visual C++ 2015 and .Net framework

3.1.2 Hardware Requirement

- PC Computers with minimum capacity
- Processor: Intel Pentium.
- RAM: 4 GB
- Storage Space: Minimum 5 Gb of free space

3.2. Problem Definition

- Work done in this field of text detection is mainly performed by using gradient features. Some of the researchers use color segmentation approach to the analysis of text.
- The only assumption that the existence of text in large gradient regions may affect some applications where the scenario is reversed.

- These algorithms only considers that pixels with large gradient values can be considered under text area.
- Various edge detection methods were presented in the literature. all are based on luminance only. Color segmentation may be required to produce good results.

3.3. OCR Applications

Optical character recognition has been performed in a numerous of applications. We discussed some of these application areas in this section.

3.3.1. Handwriting Recognition

Handwriting recognition is the capacity of a PC to get and translate intelligible handwritten data from sources, for example, paper records, photos, touch-screens and different gadgets. The picture of the written content might be detected "off line" from a bit of paper by optical scanning (optical character recognition) or clever word recognition. On the other hand, the developments of the pen tip may be detected "on line", for instance by a pen-based PC screen surface.

3.3.2. Receipt Imaging

Receipt imaging is broadly utilized as a part of numerous organizations applications to monitor financial records and keep accumulation of payments from heaping up. In government offices and autonomous organizations, OCR simplifies information gathering and analysis, among different procedures.

3.3.3. Legal Industry

Legal industry is likewise one of the recipients of the OCR innovation. OCR is utilized to digitize documents, and to specifically enter into PC database. Legitimate experts can further search documents required from tremendous databases by basically writing a few keywords.

3.3.4. Banking

Another imperative use of OCR is in banking, where it is utilized to process cheques without human intervention. A cheque can be embedded with a machine where the framework filters the sum to be issued and the right measure of cash is exchanged. This 248 | IJAMEC, 2016, 4(Special Issue), 244–249 This journal is © Advanced Technology & Science 2013 innovation has been idealized for printed cheque, and is genuinely precise for handwritten checks diminishing the hold-up time in banks.

3.3.5. Healthcare

To process printed material, medicinal services have likewise seen an expansion in the utilization of OCR innovation. Medicinal service experts continuously need to manage extensive volumes of documents for each patient, including protection frames and in addition general health forms. To stay aware of every one of this data, it is valuable to input relevant information into an electronic database. With OCR processing tools, we can extract data from structures and put it into databases, so that each patient's information is quickly recorded and retrieved when needed in future.

3.3.6. Captcha

A CAPTCHA is a system that can create and grade tests that human can pass yet current software technology can't. Malicious programmer can make software to misuse personal information on websites. Dictionary attack is assault against secret word confirmed frameworks where a programmer composes a system to over and over attempt distinctive passwords like from dictionary of most regular passwords. In CAPTCHA, a picture comprising an arrangement of letters and numbers is produced with variety of size and textual styles, distracting backgrounds, arbitrary portions, highlights and noise so that text cannot be read via OCR. Current OCR frameworks can be utilized to evacuate the noise and portion the picture to make the picture tractable by such malicious users.

3.3.7. Automatic Number Plate Recognition

Automatic number plate recognition is utilized as a mass observation method making utilization of optical character recognition on pictures to recognize vehicle registration plates. ANPR has additionally been made to store the pictures caught by the cameras including the numbers caught from license plate. ANPR innovation own to plate variety from place to place as it is an area particular innovation. They are utilized by different police forces and as a technique for electronic toll accumulation on payper-use streets.

3.3.8. ATMA

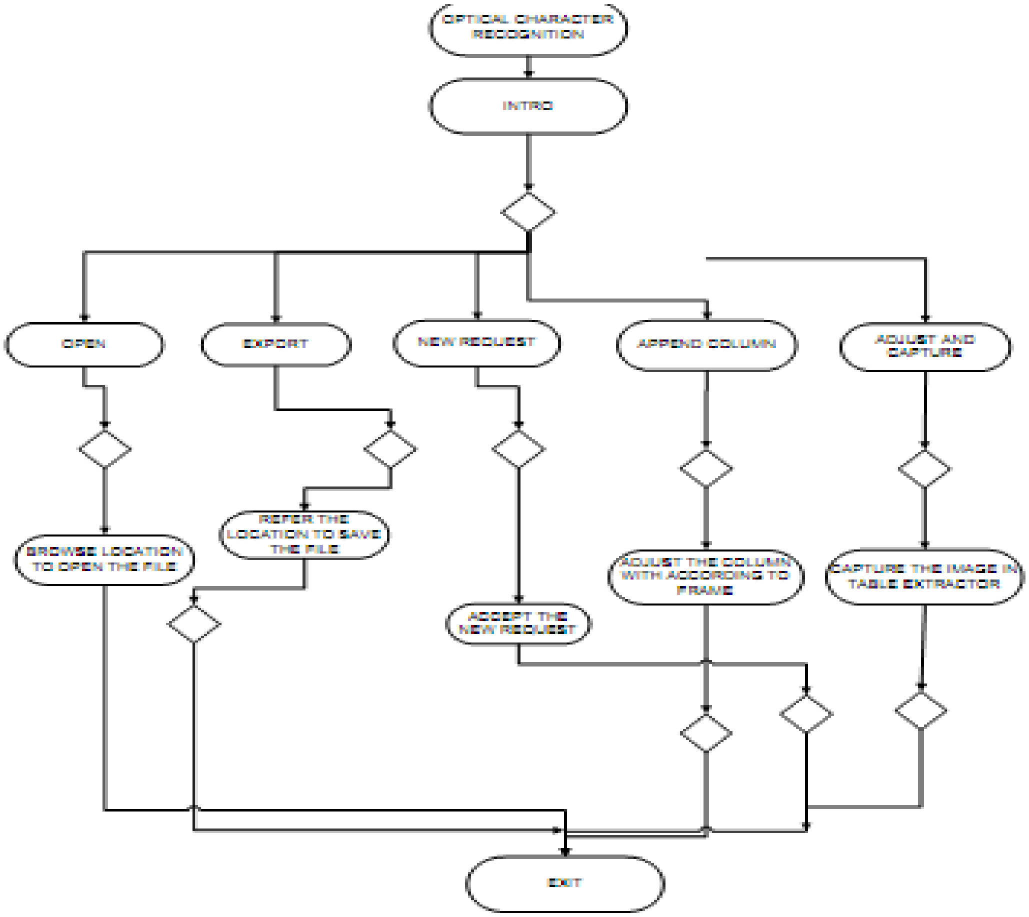
Android travel mate application ATMA: android travel mate application proposed by Mishra, Nitin, and C. Patvardhan, that It empowers Tourists and Travellers to effortlessly catch the native signboards, nation dialect Books pages, hotel menus, banners and so on. Unicode text format was obtained from content embedded in the caught image by an implicit OCR. With the goal that Travellers can translate the native Dialect Unicode content into their own nation dialect, it likewise gives translation feature.

4.1 System Design

Systems design is the process of defining the architecture, modules, interfaces, and data for a system to satisfy specified requirements. Systems design could be seen as the application of systems theory to product development. There is some overlap with the disciplines of systems analysis, systems architecture and systems engineering

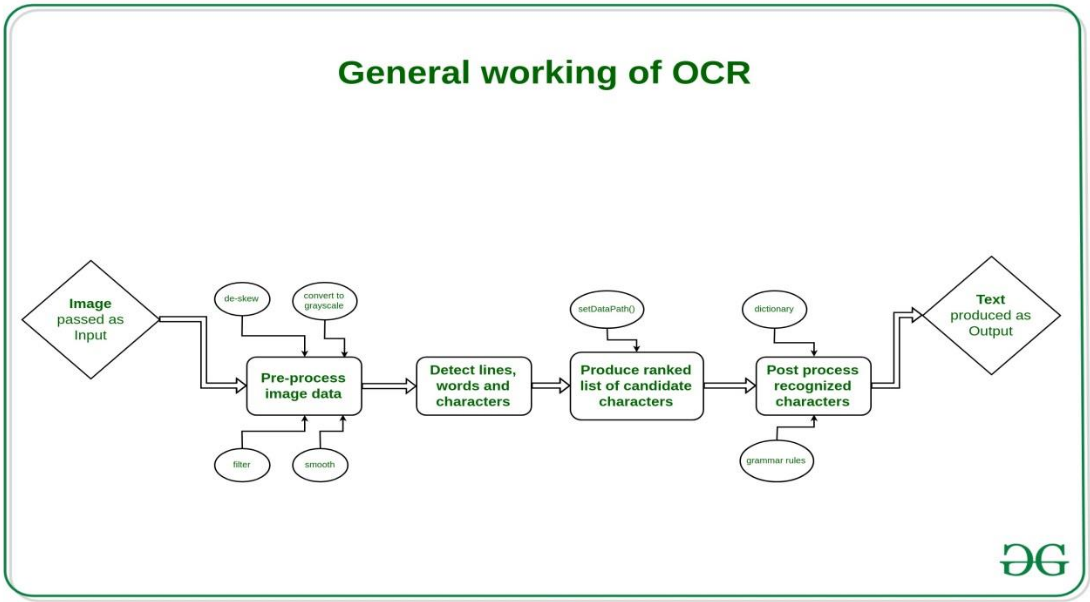
4.1.1 Activity Diagram

An activity diagram is a behavioral diagram i.e. it depicts the behavior of a system.
This activity diagram portrays the control flow from a start point to a finish point showing the various decision paths that exist while the activity is being executed in OCR Data Extraction Process.



OPTICAL CHARACTER RECOGNITION
INTRO
OPEN
EXPORT
NEW REQUEST
APPEND COLUMN
ADJUST AND CAPTURE
REFER THE LOCATION TO SAVE THE FILE
BROWSE LOCATION TO OPEN THE FILE
ADJUST THE COLUMN WITH ACCORDING TO FRAME
CAPTURE THE IMAGE IN TABLE EXTRACTOR
ACCEPT THE NEW REQUEST
EXIT

4.1.2. General Flow Diagram



Generally OCR works as follows:

1. Pre-process image data, for example: convert to gray scale, smooth, de-skew, filter.
2. Detect lines, words and characters.
3. Produce ranked list of candidate characters based on trained data set. (here the setDataPath() method is used for setting path of trainer data)
4. Post process recognized characters, choose best characters based on confidence from previous step and language data. Language data includes dictionary, grammar rules, etc.

4.2.Design of OCR

Various approaches used for the design of OCR systems are discussed below:

Matrix Matching [6]: Matrix Matching converts each character into a pattern within a matrix, and then compares the pattern with an index of known characters. Its recognition is strongest on monotype and uniform single column pages.

Fuzzy Logic [6]: Fuzzy logic is a multi-valued logic that allows intermediate values to be defined Vol. 4, No. 6 June 2013 ISSN 2079-8407 Journal of

Emerging Trends in Computing and Information Sciences ©2009-2013 CIS Journal. All rights reserved. http://www.cisjournal.org 546 between conventional evaluations like yes/no, true/false, black/ white etc. An attempt is made to attribute a more human-like way of logical thinking in the programming of computers. Fuzzy logic is used when answers do not have a distinct true or false value and there is uncertainly involved.

Feature Extraction [3]-[6]: This method defines each character by the presence or absence of key features, including height, width, density, loops, lines, stems and other character traits. Feature extraction is a perfect approach for OCR of magazines, laser print and high quality images.

Structural Analysis [6]: Structural Analysis identifies characters by examining their sub features- shape of the image, sub-vertical and horizontal histograms. Its character repair capability is great for low quality text and newsprints.

Neural Networks [6]: This strategy simulates the way the human neural system works. It samples the pixels in each image and matches them to a known index of character pixel patterns. The ability to recognize characters through abstraction is great for faxed documents and damaged text. Neural networks are ideal for specific types of problems, such as processing stock market data or finding trends in graphical patterns.

Summary

The scheming of my system has given the in general and by and large all the essential details of the submission” Optical Character Recognition”. From this any laymen user can also appreciate the design of the system. The universal information about the basic functionality of the scheme and how it works along with the user physical has been given in the completion phase.