

## Secured Digital Data Extraction of enterprise invoices images using Fuzzy Matching Techniques

R. Priyadarshini\*, Swetha<sup>2</sup> S. B, R. Kiran<sup>3</sup>, Mohammed Azeem.M. A<sup>4</sup>,  
N. Rajendran<sup>5</sup>

Department of Information Technology, B.S. Abdur Rahman Crescent Institute of  
Science and Technology <sup>1,2,3,4,5</sup>

<sup>1</sup>[rspdarshini@gmail.com](mailto:rspdarshini@gmail.com), <sup>2</sup>[kiranravichandran.kr@gmail.com](mailto:kiranravichandran.kr@gmail.com),

<sup>3</sup>[sweswethasubramanian1999@gmail.com](mailto:sweswethasubramanian1999@gmail.com), <sup>4</sup>[mohamedazeem24@gmail.com](mailto:mohamedazeem24@gmail.com)

<sup>5</sup>[rajendran.n81@gmail.com](mailto:rajendran.n81@gmail.com)

### Abstract

NLP is one of the components of Machine Learning. Topic Modeling is a subcomponent of information retrieval Information Retrieval is a broad area of research in Natural Language Processing (NLP). Image Processing is the extraction of meaningful information mainly from digitally stored images by means of by a number of processing techniques, each however may be useful for a small range of tasks as there still aren't any known methods of image analysis that are generic enough for wide ranges of tasks, compared to the abilities of a human's image analyzing capabilities. However, there are some instabilities in retrieving useful information from the Existing system which is involved with a lot of manual operations. Managing and maintaining the existing system was difficult. Like managing the invoice of all the employees which are submitting by the employees for reimbursement and making the information entry manually is tedious job and consumes lot of time. In order to overcome the issues, in this project a Secured Digital Data Extraction of Enterprise Invoice is proposed. To solve the existing discrepancies, the proposed system takes on Optical character recognition and Line deduction in Image enhancement techniques for categorization of invoices. To overcome the instabilities and poor data retrieval rate in existing system, fuzzy matching techniques and key word-based searching is proposed. The Generation and Integration phase will be combined to give the summarized legal text documents that prevent from instability and performance issues. To perform efficient segmentation of data from the enhanced image by using Line detection and Keyword search engine algorithm. Line detection algorithm HOCR Box Method and HOCR Line method uses the lattice point edge detection to extract text from images bounding box-based approach based on white spaces, case sensitive and font type as Sans-serif and Serif.TXT method will extract the text characters from the images to text file as output. By Defining the keywords and training the keywords by using Long Short Term Memory (LSTM)+ Neural network to make it as meta keyword which is used as parameters for fuzzy matching. It uses the RNN (Recurrent Neural Network) to train each time based on the invoice details that has been extracted to the text file. The semantic keyword is the parameters which is given as the fuzzy matching co-ordinates to locate the entire file by reading line by line and check for the best match. It uses Neural network to train each time, based upon the input and text which has been generated from the invoice images to get the exact match of the details to the CSV file. Then the CSV file is converted in to the excel file for detailed view of invoice information. Due to proposed algorithm and its comparison of accuracy – compared to old systems, drawbacks – rectify, the proposed system has increased from 95 – 98% accuracy – recall and precision.

**Keywords:** NLP, CV2, RNN, image enhancement Fuzzy Logic and LSTM.

### 1. Introduction

The Automated Invoice Summarization model aims at designing the invoice claiming and reimbursement claiming system without manually intervention. Employees will be able to

upload the Invoice statements to claim the expenses. This proposed system is an end-to-end solution for analyzing the invoice and extracting the data from it. It will identify the invoice documents and will increase the accuracy of analyzing the invoices by extracting the text from it. Automated invoice processing can cast off many complex activities along with classification, validation and manual facts entry, as a result greatly reducing the processing time.

This proposed model works based on Optical character recognition and Line detection by using techniques of Image enhancement for classifying the invoices. Fuzzy matching techniques and Key word based searching are used to correct the irregularity in data retrieval rate. The system also works best in segmentation of data from the enhanced invoice image by employing the Line detection algorithm. Thus, Automated Invoice System makes it easier for both the employees and the admin team to claim and process the reimbursement without any delay.

## **1.2. Information retrieval (IR)**

Information retrieval System objectives in designing and imposing systems to be able to capable of offer speedy and effective content primarily based get entry to to large amount of statistics along with text, pics and so on. The facts retrieval device should interpret the contents of the document in a set and rank them consistent with the relevance to the want of user.

Information retrieval is the study of scanning for data in a record, finding out archives themselves, and furthermore drilling down meta-data that portray information such as text, audio or images. The IR system will respond by retrieving the relevant output, in the form of documents, about the required information and will return the required documents related to the desired information.

Natural language processing (NLP) is an automatic and intelligent analysis of text data and meaningful insights can be derived from the analyzed data. Some of the popular NLP techniques includes Text analysis, Sentimental Analysis, Information Extraction and Information Retrieval (IR). The objective of NLP is to peruse, interpret, comprehend, and perceive human languages in a significant manner.

The NLP mainly depend on Artificial Intelligence to extract context from human. The applications of Natural Language Processing are difficult since human in an exact, unambiguous, very organized programming language. Natural language processing allows computer systems talk with humans in their personal language and scales other language-related tasks. NLP makes it possible for computers to study text, listen speech, interpret it, degree sentiment and decide which components are important. NLP tasks breaks down language into shorter, elemental pieces, try to recognize relationships among the portions and discover how the portions work collectively to create meaning.

## **1.3. Fuzzy Matching**

A Fuzzy matching is a shape of computer-aided translation, or CAT, and may be used to healthy sentences or sections of text to be translated. Fuzzy matching is a process that produces a result based totally on word-based totally matching keyword to extract information which is green than conventional database searching. Fuzzy matching is a method that provides an improved ability to process word-based matching queries to find matching phrases or sentences from a database. Fuzzy string matching is a form of search as a way to locate fits even when users misspell words or enter only partial words for the search. It is also recognized as approximate string matching..

The technique mainly deals with process of finding string that approximately matches the given text. Once an exact match is not found for a sentence or phrase, fuzzy matching can be applied. Fuzzy matching allows you to identify non-exact matches of your target item. Fuzzy matching attempts to find a match which, even though no longer a a hundred percent suit, is above the brink matching percent set by using the application

#### **1.4. Image Enhancement**

Image Enhancement is necessary to restore or improve the quality of a tainted image by improving the quality of image files, the OCR output is more accurate and less likely to have errors. In image processing, line detection is an algorithm that takes a group of  $n$  edge points and reveals all the lines on which these edge factors lie. Image enhancement is the procedure of improving the fine and information content material of original records before processing.

Image Enhancement process belongs to image pre-processing methods by improving the contrast, sharpening the image so that it is better suited for further processing or analysis. These methods are based on subjective image quality criteria. Line detection is a method that is used in image processing to detect any shape, if that shape can be represented in mathematical form. It can detect the shape even if it is broken or distorted a little bit. There are three methods of Image enhancement such as Point processing, Spatial filtering and Image coloring. The point processing includes enhancement of contrast, modelling the histogram and averaging the image. The spatial filtering constituents Linear filters, Non-Linear filters, Edge detection and Zooming of the images. The image coloring consists of Pseudo coloring and False coloring.

#### **1.5. Optical Character Recognition**

Optical Character Recognition (OCR) is a technology that permits the digitization of scanned snap shots with printed or handwritten text into machine-readable facts which could later be used for digital editing. Tesseract OCR is an optical character recognition engine for various operating systems. OCR offers a massive saving of storage space by compacting paper documents into electronic documents, search ability is vastly improved for printed texts Tesseract OCR extracts text from images and documents without a text layer and outputs the document into a new searchable text file, PDF, or most other popular formats. Tesseract is highly customizable and can operate using most languages, including multilingual documents and vertical text.

This technology is used as text recognition by distinguishing printed or handwritten text characters inside digital images of physical documents such as scanned paper document. This process involves in examining the content of document and translating the characters which is used for processing the data.

#### **1.6. Problem Definition**

In Existing system, it involves lot of manual operations. Managing and maintaining the existing system is difficult. Like managing the invoice of all the employees which are submitting by the employees for reimbursement and making the information entry manually is tedious job and consumes lot of time. The employees have to manually go and submit their expense reports to the AWS team (Admin Work Space). The AWS team then manually enter the details in the system and checks for any discrepancy and process the reimbursement. This makes it difficult for both the employee's and the AWS team.

The main issue of manual information entry of invoices is, there are chances of human error which may leads to false reimbursement. Traditional database search of invoice information increases the retrieval time. Maintaining the number of invoices of employee in an enterprise is tedious job when it comes to huge numbers of reimbursement request.

In Need for the customizable image to text retrieval system, the existing image to text retrieval using bounding box techniques there are lot of issues retrieval of information not in proper format like extract all the text from the images as whole not in an structured manner will be tedious job to make as an useful information for reimbursement of invoices and less accurate retrieval of text from the invoice images and these kind of systems will not use semantic word based retrieval. Instead the existing automatic conversion will not fetch the best results among the conversions done. Therefore, there is need for the automated invoice summarization model.

## 2. Related Work

Cynthia Pitou and Jean Diatta”,[2019].Textual Information Extraction in Document Images Guided by a Concept Lattice. This paper has proposed that the concept of extracting the relevant data from a set of document images with the use of Text Information Recognition. This involves identifying and understanding text in document images which helps in determining where to locate it, and where to translate into plain text. Regions of prototype will be derived from lattice concepts of a formal framework are developed from partition clusters. Text processing is a processing of meaningful human data into an intelligent format for a computer program. Document processing involves analysing and understanding the document. This literature provides solutions for textual information in an understandable format by means of deep learning and text analysis. This model does not consider a wide variety of document relationships. It does not deal with various latent relationships in the document.

“P.Saravana Kumar, T.V.P.Sundararajan, J Poornimasre”,[2019]. Fuzzy Based Estimation of Enhanced color Illumination for Digital Images. This paper proposed that the concept of Social Interactions among living beings is more critical with the fundamentals of digital forgery, so that a step by step procedure for image detection can be easily manipulated by editing software using various media elementary image representation of any information. Detection of image forgery is a tool for detecting unlicensed image processes using the fuzzy classification instead of SVM classifier for precise tests. There are several characteristics of digital images, the colour and luminosity of each pixel and the resolution and format of each image. Fuzzy matching results in high accuracy rate in information retrieval. On the other hand, it is difficult to identify strangled images.

“C.Sugapriya”[2018]. Quality Improvement of Image Processing Using Fuzzy Logic System. This paper has proposed a concept that the transmission and reception of digital grey image is achieved by various digital products which is often decreased by pulse noise. Image processing consist of two fields namely noise reduction and data security. The essential forms of noise include various types such as Impulse noise, multiplicative noise and additive noise. Denoising is one of the main tasks in digital image processing and pre-processing phase. The merit of this methodology is that it retrieves the information from renders the objects recognizable and obscures images. For the reduction of impulses noise, there are several median filters available with these methods which must be improved. However, the photo denoising is insufficient and takes too much time.

“Angamuthu Thilagavathy, Arulgnanaprakasam chilambuchelvan”,[2018].An Improved Fuzzy Based algorithm for Detecting Text from Images using Stroke Width Transform. Photo processing edge detection is a major activity in the area of unique processing technology. Image processing and soft computing are two efficient techniques. Soft computing is a growing is a growing field of confusion trading in tribulations and for detecting the image edges fuzzy logic is used. The simple detection of the canny border is typically used for SWT text detection. First of all, image is divided into parts with a binary template. In the fluctuation point, the image input has been translated to the [0 1] domain. The merits are it excels in automatic text classification tasks and this approach is Highly effective and feasible. However, this approach has less run time efficiency.

“K.N.Natei,J.Viradiya,S.Sasikumar”,[2018]. Extracting Text from Image Document and Displaying Its Related Information. These image texts are highly available nowadays and they are very important in representing, describing and transferring information which help peoples in communication, solving problems, availability, creation of new types of jobs, cost effectiveness, productivity, globalization and cultural gap. These image texts are highly available nowadays and they are very important in representing, describing and transferring information which help peoples in communication, solving problems, availability, creation of new types of jobs, cost effectiveness, productivity, globalization and integrated MATLAB code with MATLAB/Simulink tool and the proposed system is tested by Digital Image Binarization Competition (DIBCO) Connected Component, Edge Based, Text Extraction, Image Text. The Advantages are Extraction of text from image documents is very important in different areas

nowadays. The proposed algorithm gives good performance in text extraction by combining two algorithms, Edge Based and Connected Components. The extracted text recognition done by OCR with better accuracy and finally audio output produced. On the other hand, the paper does not include handwritten and complex font text.

“Aruna Bajpai, Virendra Singh Kushwah”, [2019]. Importance of Fuzzy Logic and Application Areas in Engineering Research. This paper has proposed about the fuzzy logic applications and finding the recent contributions by using the fuzzy theory and their computational ability. In addition, this paper also includes taxonomy and provides the proposal of utilizing the technology in text mining domain as an application for future design and development. The fuzzy logic is rich domain of applications. The ability of fuzzy system makes it more and more acceptable in our real-world problem-solving technique. The potential of fuzzy logic makes it more valuable for their applicability. The problem-solving technique is much transparent and flexible to scale and minimize the requirements according to the application needs. This paper also describes that the benefits of the fuzzy logic is more flexible and generality in the formulation and solution of problems. The merits are it excels in rapid operations and decision control and it has high precision on outcomes. However, this approach lack real time response and it is not advisable for large amount of data.

“Karez Abdulwahhab Hamad, Mehmet Kaya”, [2016]. A Detailed Analysis of Optical Character Recognition Technology. This paper elaborates on the challenges of the Optical Character Recognition, important phases of architecture which includes pre-processing, segmentation, normalization, future extraction, classification and post processing. The performance and accuracy of OCR is directly dependent upon the quality of input documents. 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. The high accuracy in character recognition can be obtained only if each phase of the architecture is implemented. The merits are it is Highly effective and feasible. However, the OCR degrades the desired characteristics of the images causing poor segmentation of the image is taken in natural environment with uneven shadows and lighting.

## **4. Methodology**

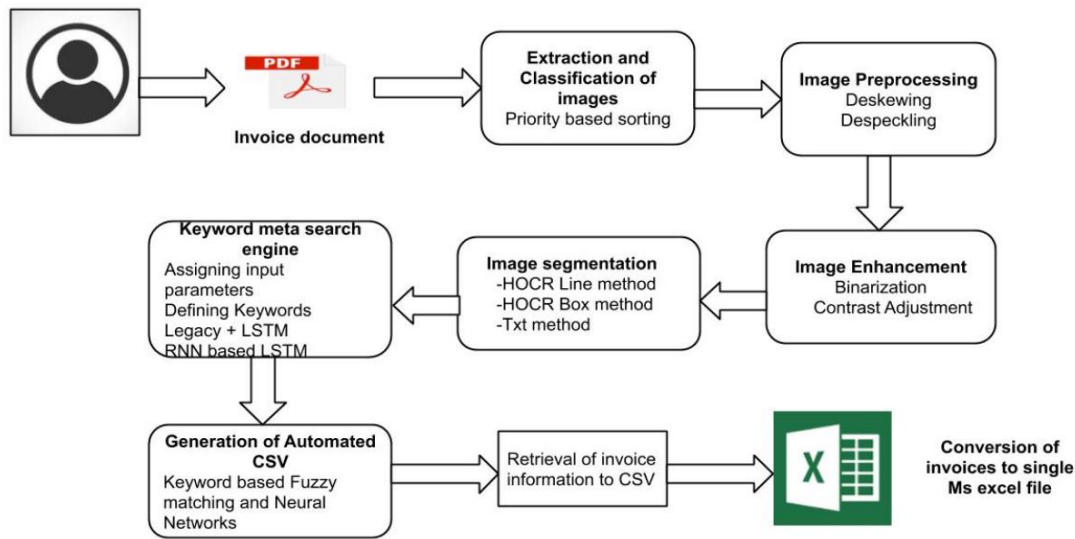
### **4.1. Proposed System**

To solve the existing discrepancies, the proposed system takes on Advanced Optical character recognition and Line detection in Image enhancement techniques for categorization of invoices. To overcome the instabilities and poor data retrieval rate in existing system, fuzzy matching techniques and key word-based searching is proposed. The processing of the images are carried on with the image enhancement techniques, along with the RNN based LSTM (Recurrent Neural Network and Long Short term Memory) is used to train keywords in the keyword based meta-search engine. Semantic keywords which are generated by the key word based meta search engine for implementing fuzzy matching technique to extract information from the extracted text from the invoice images.

The Generation and Integration phase will be combined to give the summarized text documents that prevent from instability and performance issues. To perform efficient segmentation of data from the enhanced image by using Line detection and Keyword search engine algorithm. The proposed model improves significantly the extraction rate of an invoice information by using the fuzzy matching techniques to find the best matches based the keywords based on the keyword invoice repository.

Significant topics produced over multiple documents of same scenario over the invoices and categorize the invoice documents with the manifest accuracy using the fuzzy matching techniques. It will help the enterprises in a productive way with the end goal such as:

- Reduces processing time.
- Accurate method of Invoice information retrieval.
- High Accuracy in invoice information from the invoice images.



**Figure .1.1 Secured data retrieval fuzzy matching based system**

In a corporate company, there will be a greater number of employees who claims their reimbursement money in the Admin Workspace (AWS) team by submitting the spending during business travel expenses in the form of invoices to reimbursement the claim amount. It is tedious job for AWS team to manually enter the invoice information of each employee and try to initiate the reimbursement process is not an easy task. This system is used to extract the useful information from the invoices of the employee to claim the reimbursement amount. The Employee must submit the invoices in the form of PDF document to the AWS Team.

Once the invoice is submitted as PDF the system will extract each and every invoice in the pdf file based on priority-based sorting and extract each invoice as an image in an JPEG format. These images are further De-Speckled to reduce the non - uniform alignment of pixels called Noise. Speckle noise are common form of noise in digital images which are undesired fluctuation of color and Luminance. The main causes of Speckle noise are electricity, heat and illumination of sensor level. Once the process of De-skewing and De-speckling is completed, it results in Pre-processed Image.

Now the pre-processed Raw image is further enhanced by binarization and contrast adjustment to make the clarity of image to extract the information in a more accurate manner.

Further the enhanced images are further processed by segmentation of images in order to extract the invoice information by using HOOCR Line method, HOOCR Box method and TXT file. HOOCR Line Method uses the lattice point edge detection to extract text from images. HOOCR Box Method uses bounding box based approach based on white spaces, case sensitive and font type as Sans-serif and Serif.TXT method will extract the text characters from the images and provide text file as output. using Long Short -Term Memory (LSTM), an Artificial recurring Neural network in the field of deep learning. This Neural network architecture further make it as meta keyword which is used as parameters for fuzzy matching.

**Table 1 : Keywords Extracted from the Text file of Enhanced Invoice images**

| S.NO | Keywords | Keywords Description                                                                      |
|------|----------|-------------------------------------------------------------------------------------------|
| 1.   | Name     | Name of the customer                                                                      |
| 2.   | Product  | Product Description of the customer's Purchase say HP Laptop 16.9 Inch Windows 10, 500GB. |
| 3.   | Price    | Amount of the product in INR 53,234                                                       |

|    |              |                                                                       |
|----|--------------|-----------------------------------------------------------------------|
| 4. | Invoice date | Date of the invoice which has been generated.                         |
| 5. | Invoice no   | Invoice number which is unique for each order placed by the Customer. |

**Table 2 : Invoice Information Extracted to the Excel file**

| Filename                         | Name  | Product                                   | Price  | Invoice date | Invoice no              |
|----------------------------------|-------|-------------------------------------------|--------|--------------|-------------------------|
| OD1088336<br>7243426200<br>0.txt | Kiran | KC Fab Casuals Blue<br>Shoes              | 500.0  | 06-04-2017   | FOU1L1DL18-<br>00000017 |
| OD1108208<br>7883965200<br>0.txt | Kiran | Nokia 105 Dual Sim<br>2017                | 1380.0 | 23-11-2017   | FABAW41800007139        |
| OD1109469<br>6688891800<br>0.txt | Kiran | Headphones Flipkart<br>SmartBuy Wired     | 339.0  | 08-12-2017   | B1804754601             |
| OD1156331<br>8222868300<br>0.txt | Kiran | SK 101A IIK<br>COLLECTION<br>Analog Watch | 2271.0 | 07-02-2017   | 4832859865901100        |
| OD1142643<br>1489750200<br>0.txt | Kiran | Knotyy Beanie Cap                         | 399.0  | 26-12-2018   | APFAR4QQXFZGGE          |

It uses the RNN (Recurrent Neural Network) to train each time based on the invoice details that has been extracted to the text file. Semantic keywords for invoice have been generated for implementing fuzzy matching technique to extract information from the text file in order to make it as a semantic keyword. The table 1 shows the keyword extracted from the text file of the extracted invoice images.

The semantic keyword is the parameters which is given as the fuzzy matching co-ordinates to locate the entire file by reading line by line and check for the best match the text file. It uses Neural network to train each time, based upon the input and text which has been generated from the invoice images to get the exact match of the details to the CSV file.

Then the CSV file is converted in to the excel file for detailed view of invoice information.

## 6. Proposed Modules Description

The following modules are explained in the following section,

- Extraction, Classification of Invoice Images.
- Image Pre- Processing.
- Image Enhancement.
- Image Segmentation.
- Defining Keyword Meta Search Engine.

- Generation of Automated CSV, Excel File.

### 6.1 Extraction, classification of invoice images

The Invoices of the employees for distinct claims are considered. These invoices are categorized as manual or automated based on which they are generated. These invoices converted into Portable document format (.PDF) using document scanner software like Cam Scanner, Adobe Scanner etc by recognizing the characters in the invoice and enhancing the partially scanned characters. By Using Open Source Computer Vision Library (CV2) scanned PDF documents are converted into Joint Photographic Expert Group (.JPEG) for each invoices.

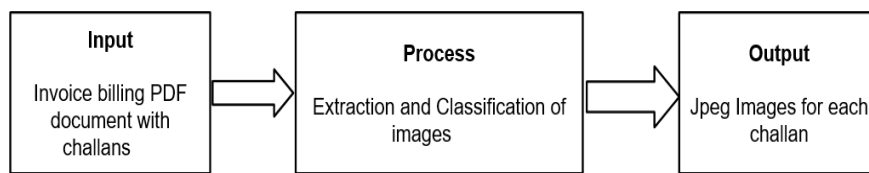


Figure 2 - Flow of Data Extraction

### 6.2 Image Pre-processing

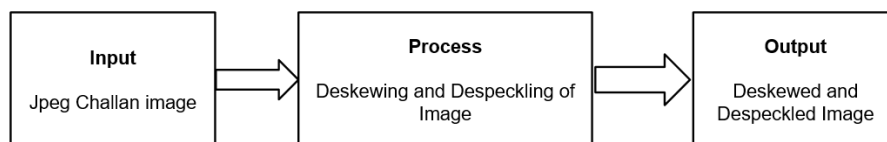


Figure 3 Flow of Image Pre-processing

The JPEG Images extracted from the PDF documents are now pre-processed by Open CV for De-Skewing an Image by converting it into Grey Scale, edges and rotating by the detected skew angle. These images are further De-Speckled to reduce the non - uniform alignment of pixels called Noise. Speckle noise are common form of noise in digital images which are undesired fluctuation of colour and Luminance. The main causes of Speckle noise are electricity, heat and illumination of sensor level. Once the process of De-skewing and De-speckling is completed, it results in Cleaned Image.

### 6.3 Image enhancement

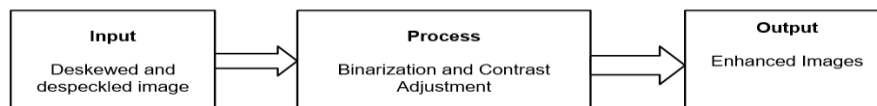


Figure 4 Flow of Image Enhancement

The Cleaned images that has uniform alignment of pixels in the detected angle are considered. These images are subjected to Binarization process which is converting the grey scale images into Black and White (i.e) Binary Image by re-building the features of the pixel from the Multi-toned images into 2 toned images. The Binary Images are further enhanced by changing the contrast. Contrast of an image is defined as difference in the luminance (Color) that makes an image recognizable, thus resulting in the form of enhanced invoice images.

### 6.4 Image segmentation

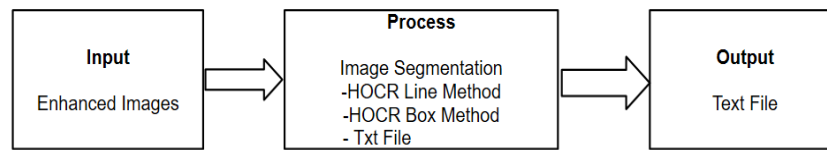


Figure 5 Flow of Image Segmentation

The Enhanced images are now further processed by segmentation of images in order to extract the invoice information by using HOOCR Line method, HOOCR Box method and TXT file. HOOCR Line Method uses the lattice point edge detection to extract text from images. HOOCR Box Method uses bounding box based approach based on white spaces, case sensitive and font type as Sans-serif and Serif. TXT method will extract the text characters from the images and provide text file as output.

**Algorithm 1 :Tri-level Segmentation based HOOCR Line :**

Input : Enhanced images

Key Parameters: Enhanced image, edge height and edge Width.

Process : HOOCR Line method -edge detection

Output: extraction of Text from Enhanced images

Algorithm:

Step 1 : Begin the process

Step 2: Input decide the range of  $\rho$  and  $\theta$  (Rho and Theta).

Step 3: Create a 2D array called the accumulator representing the dimension (*num\_rhos*, *num\_thetas*)

Step 4: Perform edge detection on the original image.

Step 5: For every pixel on the edge image, check whether the pixel is an edge pixel.

Step 6: Edge pixel will be loop through all possible values of  $\theta$ .

Step 7: Calculate the corresponding  $\rho$ , find the  $\theta$  and  $\rho$  index in the accumulator, and increment the accumulator base on those index pairs.

Step 8: End the Process.

**Algorithm 2 : Trilevel Segmentation based HOOCR Box Algorithm:**

Input : Enhanced images

Key Parameters : Image.PNG,

H - Height of the image

W - Width of the image.

Process : HOOCR Box method -edge detection

Output: extraction of Text from Enhanced images.

Algorithm :

Step 1: Begin the process

Step 2: Input as the Image and read the image using `cv2.imread(Filename)`.

Step 3: Assign the height and Width (*h,w*) of the image

Step 4: By using the box method split the images in to Rectangle box by using `pytesseract.image_to_Box (image)`.

Step 5: Using the `cv2.rectangle` give the parameters of the image in VIBGYOR image range from 0 to 255.

Step 6: End the process.

The above algorithms enhance and recognizes the images and the characters in the same.

### 6.5 Defining key word meta search engine:

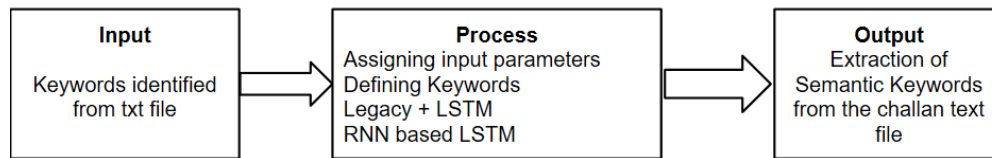


Figure 6- Flow of Defining Key Word Meta Search Engine

The keywords from the text file are identified, defined and trained by using Long Short Term Memory (LSTM), an Artificial Recurring Neural network in the field of deep learning. This Neural network architecture further make it as meta keyword which is used as parameters for fuzzy matching. It uses the RNN (Recurrent Neural Network) to train each time based on the invoice details that has been extracted to the text file. Semantic keywords for invoice has been generated for implementing fuzzy matching technique to extract information from the text file in order to make it as a semantic keyword.

#### Algorithm 3 : Semantic keyword based LSTM Model

Input : semantic keywords from text document

Key Parameters:

Prev\_Ct - Previous Cell state.

Ht - Hidden state.

It - Input Gate.

Ot = output gate

Ct = Candidate

Process: Uses the memory cell to train the model.

Output : Defines the semantic keyword for parameters to Fuzzy matching.

Algorithm:

Step 1: Begin the process.

Step 2: Define the Previous cell state and hidden state and input as text extracted from the image as key word to be trained.

Step 3: Then Combine the input and hidden state combine = prev\_ht + input.

Step 4: The candidate Holds viable values to feature to the mobile state.

Step 5: *Combine* the get's fed into the input layer. This layer decides the data from the candidate should be passed to the new cell state.

Step 6: Once the computation of forget layer, candidate layer, and the input layer, the cell state is calculated using those vectors and the previous cell state.

Step 7: End the process.

### 6.6 Generation of Automated csv, excel file

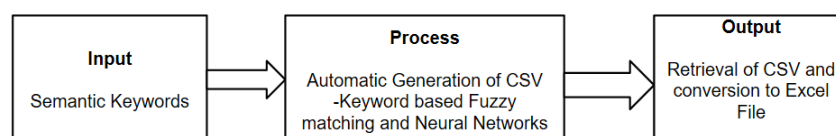


Figure 7 -Flow of Generating Automated CSV, EXCEL File

The semantic keyword is the parameters which is given as the fuzzy matching co-ordinates to locate the entire file by reading line by line and check for the best match the text file. It uses Neural network to train each time, based upon the input and text which has been generated from the invoice images to get the exact match of the details to the CSV file. Then the CSV file is converted in to the excel file for detailed view of invoice information.

**Algorithm 4 :** Fuzzy string matching based Keyword retrieval algorithm:

Input: Text in all documents

Process: Fuzzy String Matching based on keyword from the text document

Output: String with high Fuzzy Ratio.

Algorithm:

Step 1: Begin the Process.

Step 2: Input as the Extracted text from the enhanced image.

Step 3: Declare the phrases as the input parameters.

Step 4: Find the phrases the text document by line detection searches the text line by line.

Step 5: Calculate the fuzzy ratio for the phrases ,  
cur\_ratio = fuzz.ratio(phrase, words in text document)

Step 6: Define the condition for the fuzzy ratio.

if cur\_ratio > 90:

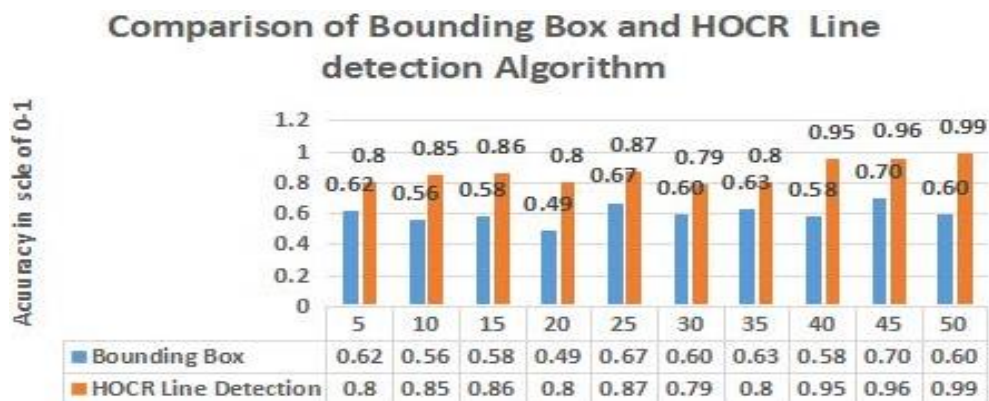
return String with high Fuzzy\_ratio

Step 7: End the Process.

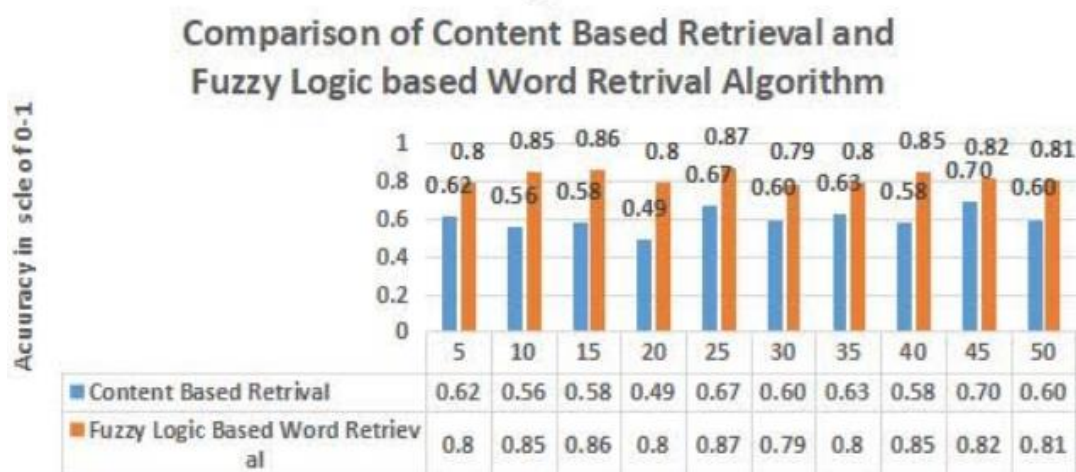
The algorithm 4 implements the fuzzy matching and converts the image to string exactly, the best priority is chosen and given as the result.

## 7. Results Discussion & Analysis

The graph depicted below in Figure 8 hows comparison between Bounding Box algorithm and HOCR Line Algorithm with X axis as “Accuracy” and Y axis as “Number of invoices tested”. In Bounding Box algorithm, Line based segmentation is done. The font specs is analyzed such as height of text layer in top, bottom and mid-level which is less efficient. And this algorithm provides accurate results only if line- height is equal to the font size whereas the proposed algorithm (i.e) HOCR Line algorithm works based on Tri-level segmentation- Line, Word and Character is used. So, the accuracy in retrieval of words is improved from the image.



**Figure 8 – Comparison of proposed HOCR algorithm with Bounding Box Algorithm**



**Figure 9 Comparison of Content Based Retrieval and Fuzzy Logic Based Word Retrieval Algorithm**

The graph depicted above shows comparison between Content Based Retrieval and Fuzzy Logic Based Word Retrieval with X axis as “Accuracy” and Y axis as “Number of invoices tested”. Accuracy is determined by number of words identified from the image divided by total number of words.

In Content Based Retrieval, the accuracy is less since the search analyses the contents of the image in place of the including keywords, tags, or descriptions associated with the picture while the proposed algorithm (i.e) Fuzzy Logic Based Word Retrieval algorithm, The accuracy is improved than the prior method content based retrieval since strings in the text are matched with best priority based on the meta-data such as keywords.

## 8. Conclusion & Future Scope

Automated Invoice System is an end-to-end solution for analyzing the invoice and extracting the data from it. The employee can upload the invoices in the system by connecting to the network. After uploading the documents, the Admin will be the one to process the reimbursement process depends upon the priority. The invoices can be of hotel bills, restaurant bills, flight tickets and cab bills. It can be of either hand-written or automated invoices. The user can view the status of the request. If the request is placed on priority, then that will be processed with importance. The system checks for discrepancy and notify the admin for that specified invoice. If the discrepancy persists, the amount will accordingly be recovered either from future claims or next salary.

The proposed model improves the rate of extraction of the text by using the fuzzy matching techniques to obtain the exact match based on keyword invoice repository. This model ensures effective segmentation of the data from the invoice image by Keyword search engine algorithm and Line detection algorithm. This approach helps the enterprises to reduce the processing time, increase the accuracy in extraction of invoice information from the images by 10% than the existing system which involves in lots of manual operations.

The application has been successfully deployed by resolving the minor issues which has been found earlier. This application is very useful to claim the expenses and to get the reimbursement. The application is serving its purposes very efficiently. It is very much user-oriented and adaptable to further future enhancements.

The present system can be extended to process invoice statements of various other languages. At present it is used only in India, can be extended to use the system across the globe. In future, we intend to extend the model by cognitive intelligence, the ability of reasoning and solving the problem. Also, the model can further be improvised by Artificial Neural Network which will be an added advantage since it contributes in improvising image processing and character recognition.

## References

- [1]. “P.Saravana Kumar, T.V.P.Sundararajan, J Poornimasre”,[2019]. Fuzzy Based Estimation of Enhanced color Illumination for Digital Images.
- [2]. “C.Sugapriya”[2018]. Quality Improvement of Image Processing Using Fuzzy Logic System.
- [3]. “Angamuthu Thilagavathy, Arulgnanaprakasam chilambuchelvan”,[2018].An Improved Fuzzy Based algorithm for Detecting Text from Images using Stroke Width Transform.
- [4]. “K.N.Natei,J.Viradiya,S.Sasikumar”,[2018]. Extracting Text from Image Document and Displaying Its Related Information.
- [5]. “Aruna Bajpai, Virendra Singh Kushwah” ,[2019].Importance of Fuzzy Logic and Application Areas in Engineering Research.
- [6]. “Karez Abdulwahhab Hamad, Mehmet Kaya”,[2016].A Detailed Analysis of Optical Character Recognition Technology.
- [7]. “Cynthia Pitou and Jean Diatta”,[2019].Textual Information Extraction in Document Images Guided by a Concept Lattice.

## 12. Copyright Forms

You must include your fully-completed, signed SERSC copyright release form when you submit your paper. WE MUST HAVE THIS FORM BEFORE YOUR PAPER CAN BE PUBLISHED IN THE JOURNAL. The copyright form is available from journal home page. Authors should send their copyright forms to FAX. +82-70-7614-3027 or E-mail [journal@sersc.org](mailto:journal@sersc.org).



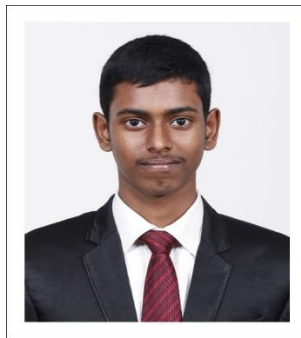
**Dr.R.Priyadarshini**, was born on 19th January 1982, in Madurai, Tamilnadu. She received her B.E. degree in Electrical and Electronics Engineering from Jerusalem College of Engineering, University of Madras, Chennai in the year 2003. She did her M.Tech. in Information Technology from College of Engineering, Guindy, Anna University, Chennai in the year 2006. She is currently pursuing her Doctor of Philosophy in the Department of Information Technology at B.S. Abdur Rahman Crescent Institute of Science and Technology. She has got more than 10 years of teaching experience in various institutions and one year of experience as software engineer at the beginning of the career. She has worked as a freelancer in web development using PHP. She has published 8 papers in peer reviewed journals, 9 papers in international conferences and 3 papers in national conferences. Her area of interest includes Semantics in Big Data analytics.



Swetha S.B is a Student at B.S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Tamil Nadu, India. She is Currently Pursuing B. Tech Degree in Information Technology. Her Area of Interest are NLP, Machine Learning and Deep Learning.



R Kiran is a Student at B.S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Tamil Nadu, India. He is Currently Pursuing B. Tech Degree in Information Technology. His Area of Interest are Big data, Machine Learning and Web Development.



Mohamed Azeem M. A is a Student at B.S. Abdur Rahman Crescent Institute of Science and Technology, Vandalur, Tamil Nadu, India. He is Currently Pursuing B. Tech Degree in Information Technology. His Area of Interest are Deep Learning, Machine Learning and Cyber Security.



N. RAJENDRAN received the B.E. degree in Computer Science and Engineering from Bharathiar University, Coimbatore, India, in 2002, the M.E. degree in Software Engineering from Anna University, Chennai, India, in 2007. He is currently pursuing his Ph.D. degree in Information Technology from B.S. Abdur Rahman Crescent Institute of Science and Technology. He has 15 years of teaching experience and he is currently working as an Assistant Professor (Senior Grade) in the Department of Information Technology at B.S. Abdur Rahman Crescent Institute of Science and Technology, Chennai, Tamil Nadu, India. His current research interests include Cyber Security, Mobile Ad hoc Networks, Big Data Analytics and Internet of Things.