IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

KYC Document Image Analysis

[1]Venkatesh Ratnaparkhe, [2]Yash Sapre, [3]Rahul Agarwal, [4]Vedant Kharkar

[1]UG Student, [2]UG Student, [3]UG Student, [4]UG Student

[1]Department of Computer Engineering,

[1]Marathwada Mitra Mandal's College of Engineering, Pune

Abstract: Now-a-day the world is moving towards digital technology, automation and image capturing devices are growing rapidly. Automation using various advanced technology helps to reduce manual effort as well as errors. OCR, an abbreviation for Optical Character Recognition, is a technology capable of minimizing human effort and eradicating errors. It is a technology that extracts text from the images or documents. It allows to edit, analyze and search the extracted data . The Objective of this paper is to summarize research that has been conducted on KYC documents to fetch the significant information out of it. In this paper we analyzed the OCR process and various text processing techniques. As OCR is being used in various fields and domains so this article serves the use of OCR to reduce manual data entry task. The proposed system will help in document digitization, reduce paper work store and manage the data efficiently.

Index Terms - Optical Character Recognition, Natural Language Processing, Regular Expression, JavaScript Object Notion, Graphical User Interface.

I. INTRODUCTION

Today, optical character recognition is playing importantrole in various areas such as business, government, educationas well as banking sector. It uses a combination of special machines, like scanners, and computer programs to do this. The machines scan the document and the software processes the scanned image to extract the text from it. This makes it easier to search, edit, or store the document digitally, and reduces the need for manual typing, which can save time and reduce errors. [1] An OCR has made the document preservation, data storage and data management process efficient and effortless. Organizations are taking benefits of OCR technique to satisfy need of data management. There are various types of documents available in different fields having fixed or varying format. Organization required documents either to validate the user or keep track of individual. For e.g., bank requires KYC (Know Your Customer) documents to ensure their customer is real and to establish customer identity. Customer needs to submit their documents to bank in physical form. This process creates theoverhead of maintaining documents in physical form, keep track of each individual ultimately it becomes difficult to retrieve the documents. This process requires lot of paper work and it will have bad impact on environment. Even if organization tries to maintain the data in digital form itrequires manual data entry which can cause the error. This isthe repetitive and time-consuming task also it reduces the response time to customer. As KYC documents have crucial details so the organization could not afford any mistake relatedto customer data.

There might be different ways to reduce manual effort and eliminate errors. The point of view is that the system needs to recognize text from documents and capture significant information using various text processing techniques. Text processing includes techniques like checking threshold value generated by OCR model, regular pattern matching and splitting the text into meaningful segments and check semantic similarity. So, this process having two parts such as text recognition and text processing

II. METHODOLOGY

This section discusses in detail the research methodologyemployed in this study shown in Figure 1. As shown here employed methodology consists of seven distinct steps [10]. Eachof these steps are further discussed in subsequent sections.

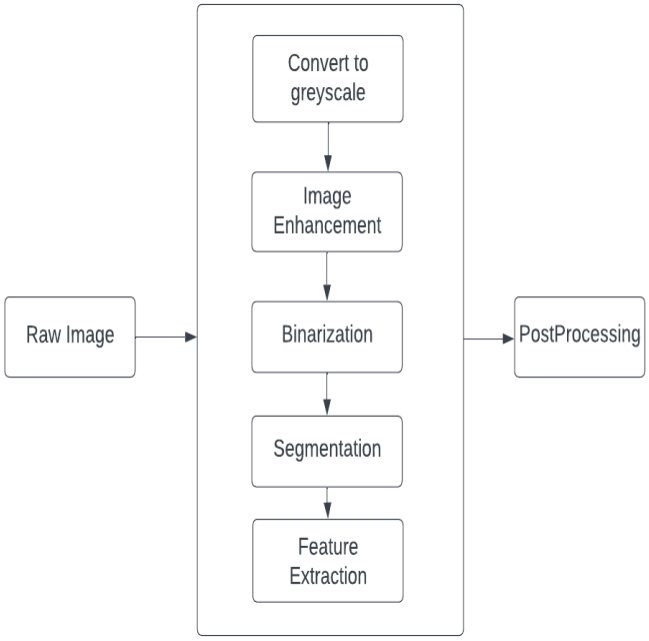

Figure 1 . Research Methodology

2.1 Convert to Greyscale image*:*

The first step in pre-processing of the image is to convert it into greyscale image. The main aim of OCR model is to recognize the text from document having a color image does not help to better read the text. It is easy to read text in black and white format. Greyscale conversion is one of the simplest methods of image enhancement. The main reason touse greyscale images instead of color images is that greyscale images simplify algorithm and reduces computational requirements. Color may introduce unnecessary information so it could be difficult to achieve good performance. There are number of methods to convert color image to greyscale image such as average method andweighted average method. Weighted method is extension of simple average method where some weight is assigned to each value. In this method center value has more weight dueto which it contributes more than other values.

$$\text{Greyscale} = 0.299*R + 0.587*G + 0.114*B \quad (1)$$

Where,

R indicates value of red

G indicates value of blue

B indicates value of blue

The weighted method is also called as luminosity method. Resulting image when applying this method is greyscale image.

2.2 Image Enhancement:

Image enhancement is the process of improving the visual quality or clarity of an image. Image enhancement can be animportant step in the process of extracting text from animage, as it can help to improve the contrast and clarity of the characters in the image and make them easier to recognize.

Some common steps involved in image enhancement forOCR may include:

a) Noise reduction:

This involves removing noise from the image that may interfere with the accuracy of the OCR model [9,12,15]. This may include techniques such as smoothing or sharpening to enhance the edges of the characters.

b) Contrast enhancement:

This involves adjusting the contrast of the image to improvethe visibility of the characters and the separation betweenthe characters and the background. This may include techniques such as contrast stretching or histogram equalization.[6]

c) Illumination correction:

This involves adjusting the overall brightness of the image to improve the visibility of the characters and the separation between the characters and the background [3]. This may include techniques such as gamma correction or histogram stretching.[4]

Image enhancement steps can have a significant impact on the accuracy of the OCR model.

2.3 Binarization:

Binarization is the method of converting the greyscale image into an image which consists of only black and white pixels. 0 indicates black color and 1 indicates white color.In this method threshold value is used to convert the pixel value into 0 and 1. Normally threshold value is 127 as it is exactly half of range 0-255. If the pixel value is greater than the threshold it is considered as white pixel else it is considered as black pixel. There are different ways to determine the threshold value. One approach is to use a fixed threshold value manually or Global thresholding technique. If the T value is constant over the image is called as Global thresholding.

Procedure for Global thresholding [15] to obtain T:

1. Select an initial estimate for T which is an average intensity of an image
2. Segment the image using T. This will produce two groups of pixels: G1 consisting of all pixels with gray level values greater than threshold and G2 consisting of pixels with values less than threshold.
3. Calculate the average gray level values denoted as μ1 and μ2 for the pixels in regions G1and G2
4. Compute the new threshold value:

$$T = ½ (μ1 + μ2) \quad (2)$$

Repeat step 2 through 4 until the difference in T in successive iterations is smaller with initial estimate.

2.4 Segmentation:

Segmentation is a necessary pre-processing step for character recognition, as it helps to isolate the characters in the image and make it easier for the OCR system to analyze and recognize them. In this method the inputs are normal images and outputs are attributes extracted from images.

Segmentation of image is done in the following sequence:

a) Line Level Segmentation
b) Word level Segmentation
c) Character level Segmentation

The process of image segmentation is commonly employed to locate and define objects as well as boundaries within images.

2.5 Feature Extraction:

In an image, features can manifest as specific structures such as points, edges, or objects. Feature extraction is the process of extracting the characteristics or features from pre-processed image that can be used to represent content of image [2]. These features may include color, texture, shape, edge information, or other visual characteristics of the characters in the image.

For example, features in PAN card are Name, PAN number, Date of birth, etc. Pattern matching is used for extracting these features. Regular expressions are used to match different patterns in raw text.

Some of the regular expressions are as follows:

a) PAN Number: “[(A-Z)]{5}[(0-9)]{4}[(A-Z)]{1}”

b) Date of Birth: “(\d{2})(/|-|1|I|i)(\d{2})/(\d{4})”

2.6 Feature recognition:

This involves using the extracted features to recognize or classify the characters in the image [5]. This may involve training a machine learning model on a large dataset oflabeled images to learn the characteristics of the different classes of characters, and then using this trained model to classify new characters based on their features.

Once the text has been extracted and transcribed into amachine-readable format [7], it can be used as a feature for various tasks such as image classification or naturallanguage processing.

2.7 Post-Processing:

Once we get the text from OCR model then we must apply various text pre-processing techniques to extract significant text or features from randomized text. For example, in Pan-card document significant features are name, dob, Father’s name and Pan-card number.

To get an exact required features with its values regular pattern matching can be used. All the features of documents have some fixed pattern. For e.g., dob is a number or string that must be in dd-mm-yyyy or mm-dd- yyyy format. Regular pattern matching is a technique used to search for patterns in each string or text. It involves using a set of rules, called a regular expression, to define thepattern that we are searching for. Regular expression is a sequence of characters that represents specific search patternwhich is to be searched. To use regular pattern matching, we first define the regular expression that represents the search pattern we are searching for. To get multiple features from text multiple search patterns could be defined. Then, we use a program or function to search for the pattern in a given string or text. If the pattern is found, the program will return a match text, such as the complete string. If the pattern is notfound, the program will return "None" or a similar value indicating that no match was found. Regular patternmatching is a powerful tool that is widely used in a variety of applications, including text processing, data validation, and data extraction. It is particularly useful for tasks that involve searching for specific patterns in large amounts of text. There might be cases where we get the text which has similar pattern as defined to find the text as required but thatwas not supposed to be considered.

For e.g., consider a pan- card document in that a regular expression (regex) to search for a full name of a person might look like this ` [A-Z]+\s[A-Z]+ `.

This regex would match any string that contains a first name and a last name, separated by a space. The first and last names are entirely composed of uppercase letters. But this pattern could also match to "GOVT OF INDIA" or "INCOME TAX DEPARTMENT" because this text also has same pattern so regular pattern matching technique is not sufficient to extract required features. Thereis one technique that can be use with pattern matching known as Token similarity. Here, Token refers to word and similarity is the degree to which the words are similar in meaning or context. This can be measured using various techniques,such as word embeddings, which are numerical representations of words that capture the relationships between words in a dataset. One way to measure tokensimilarity is to calculate the cosine similarity between the word embeddings of the two tokens. Cosine similarity is a measure of similarity between two vectors in a multi- dimensional space, and it is computed by calculating the dot product of the vectors and dividing it by the product of their magnitudes.

To calculate the cosine similarity between two tokens, we first need to represent each token as a vector.This is typically done using word embeddings, which are numerical representations of words that capture the relationships between words in a dataset. Once we have the word embeddings for the two tokens, we can calculate the cosine similarity using the following formula:

$$\text{cosine similarity} = (A * B) / (||A|| * ||B||) \quad (3)$$

Where A and B are the word embeddings for the two tokens, and ||A|| and ||B|| are the magnitudes of the vectors. The dot product (A * B) is calculated by multiplying the corresponding elements of the two vectors and summing the results. The magnitudes of the vectors are calculated by taking the square root of the sum of the squares of the elements of the vectors. A cosine similarity score of 1 indicates that the vectors are identical, while a score of 0 indicates that they are not identical.

Using the combination of regular pattern matching and Token similarity techniques required fields can be extracted from OCR generated text.

III.Evaluation :

In this section, we describe the evaluation methodology used in this work.

a) Datasets:

The main goal of this work is to improve the OCR process for KYC documents. However, extensive KYC documents are not publicly accessible due to security and privacy reasons. Hence, to check the robustness of the proposed model, it is also evaluated on manually collected pan cards.

b) Measuring Quality of OCR:

Carrasco et al. [11] present the following metrics that weused in this work. Word error rate (WER) is calculated as

$$WER = (I + S + D)/N \quad (4)$$

Where,
N represents the number of words in the ground truth text,
I the number of words inserted,
S the number of words substituted,
D the number of words deleted to get the original ground truth values.
WER values are correlated to CER values.
The Character Error Rate (CER) is defined as [13],

$$CER = (i + s + d)/n \quad (5)$$

where ,
i represents minimum character insertions,
d is the deletions to get original results,
s character substitutions to convert output to the actual text.
For example: Ground Truth: 'this is a plant'
OCR Output: 'thiz iz a plant', From the above, the CER is 0.1667%, whereas the WER is 50%. The WER value is 50% as 2 out of 4 words are correct. Therefore, WRR is 1 – 0.5 = 50% and CRR is 93.3%.

According to Carrasco et al. [11], there are threetypes of OCR errors:

i. Misspelled characters (substitutions)
ii. Lost or missing text (deletions)
iii. Spurious symbols (insertions)

Misspelled Characters (Substitutions) Error: This is an error in which OCR output has misspelled characters. Using Equation 1, this error can be calculated as,

$$s = (CER * n) - (i + d) \quad (6)$$

For example, in Figure 2, is read as "GQBPK8200C",Here "7" in the text is misinterpreted by "2".

Lost or Missing Text (Deletions) Error: is when the OCR output text misses a character. Using Equation 1, the lost character can be calculated as,

$$d = (CER * n) - (i + s) \quad (7)$$

For example, in Figure 3, is read as "AXLPA3325",Here "Q" in the text is disappeared.

Spurious Symbols (Insertions) Errors: happen if the OCR output text contains new characters inserted thatare not available in the document. Using Equation 1, the insertion character error can be calculated as,

$$i = (CER * n) - (s + d) \quad (8)$$

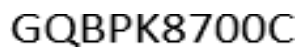

Fig. 2. Example of Misspelled Characters

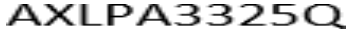

Fig. 3. Example of lost or missing text in dataset

I. IV. Result

In order to measure the effectiveness of the proposed approach, we test them on the pan card dataset.We have tested our KYC OCR system against 19 pan cards and the average WRR value we received is 0.789 and the average CRR value we received is 0.8192. Table I presents the results on pan card dataset aftercalculating CER and CRR of pan card attributes.

TABLE I. CHARACTER RECOGNITION RATE (CRR) OF ATTRIBUTESPAN CARD DATASET

File Number	*Pan number*	*Name*	*Father's Name*	*Date*
1	1	1	1	1
2	1	1	1	1
3	0.9	1	1	1
4	1	1	0	1
5	1	1	1	1
6	1	1	1	1
7	1	0	0	1
8	1	0	0	1
9	1	0.0625	0.4	1
10	0	0	0	0
11	1	1	1	1
12	1	1	1	1
13	1	1	1	1
14	1	1	1	1
15	1	1	1	1
16	1	1	0	1
17	1	1	1	1
18	0.9	0	0	1
19	1	1	1	1

Table II presents the results WRR calculated on pan carddataset. Table III presents Average CRR and Average WRR.

TABLE II. Word Recognition Rate

File Number	*Correct prediction of words*
1	4
2	4
3	3
4	3
5	1
6	1
7	2
8	2
9	2
10	0
11	4
12	4
13	4
14	4
15	4
16	3
17	4
18	1
19	4

Table III Average CRR and WRR

Document Type	Average CRR	Average WRR
Pan Card	0.8192	0.789

Bar Chart represents character recognition rate based on PAN card fields such as name, Fathers name, birth date and pan number. [8] Character recognition rate includes validating each field character by character. X-axis consists of document number where Y-axis indicates character recognition rate. If the predicted field is exactly similar to accepted field, then CRR is 1. It is expected to predict all the fields with CRR 1.

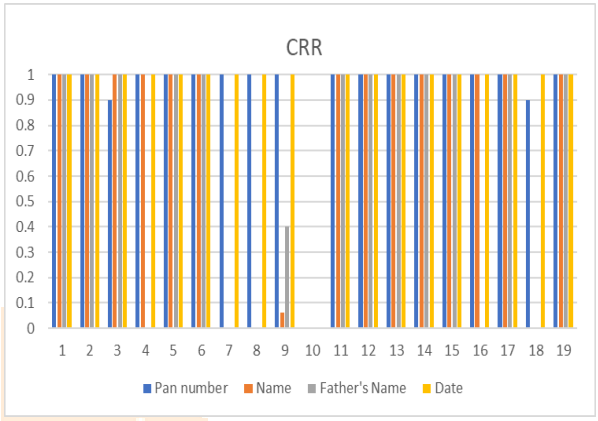

Fig. 3. Bar Chart of character recognition rate based on PAN card fields

Given below a graph is Bar chart represents Word recognition rate based on different pan card fields. X-axis consists of document number where Y-axis indicates word recognition rate. If all the identified fields are similar to expected fields then WRR will be 1. [13,14] WRR includes checking each individual field with ground truth i.e. expected field. It is expected to predict all the fields with WRR 1.

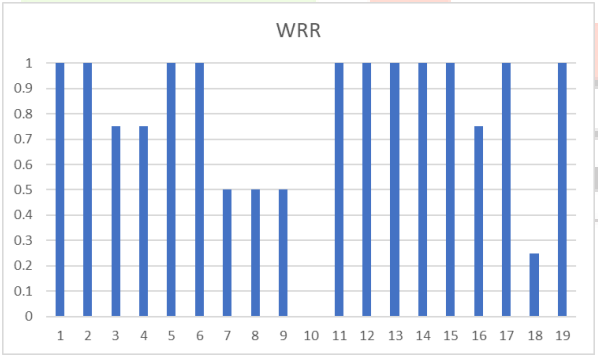

Fig. 4. Bar Chart of Word recognition rate based on PAN card fields

Input and expected output is given below, input to the system is raw or scanned image of the document given in fig.5.It is better to provide quality input image to system in order to get accurate results. Document will be processed by model to fetch the required data out of it.

Fig. 5.PAN Card document

The following result is obtained by using proposed methodology. Extracted data will be in the form of key-value pairs. Key is the attribute of the document and value indicates value of the attributes. Key-value format is readable and easy to understand.

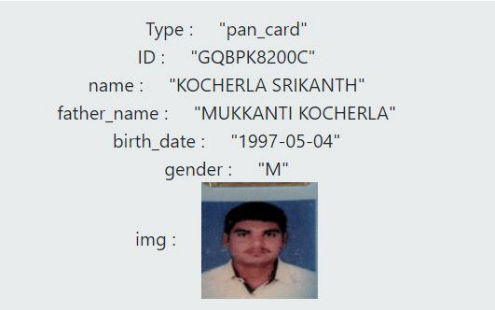

Fig. 6. Extracted PAN card data fields

V.Acknowledgemnt:

The authors would like to thank Professor, Department of CSE, Marathwada Mitra Mandal's College of Engineering for encouragement, support and guiding project.

VI Conclusion and Future Work:

Text detection, extraction and recognition is possible due to OCR technique. Amount of research is being done to improve the accuracy of OCR model. There are many languages across the globe this research paper focuses on Latin-script text detection and extraction. We have discussed OCR techniques along with Natural Language processing pipeline to detect and extract significant data from documents for example Pan card, Aadhar card, etc. Natural language processing includes various techniques to process the raw data generated by OCR model. Proposed methodology is independent of format of the documents. This methodology can be used over different format documents. In future different models can be built to perform text extraction on different documents.

Upcoming researchers are recommended to consider complexity of input and speed of process while extracting significant data out of documents.

VII. References :

[1] J. Menon, M. Sami, R. A. Khan, M. Uddin, "Handwritten Optical Character Recognition (OCR)-A Comprehensive Systematic Literature Review (SLR)", IEEE, August 2020

[2] R. Mittal, A. Garg, "Text extraction using OCR: A Systematic Review", Second International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE, 15-17 July 2020.

[3] Jie Ding, Guotao Zhao, Fang Xu, "Research on Video Text Recognition Technology Based on OCR", 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), IEEE, 10-11 February 2018.

[4] Matteo Brisinello, Ratko Grbic, Dejan Stefanovic, Robert Peckai Kovac, "Optical Character Recognition on images with colorful background", IEEE 8th International Conference on Consumer Electronics - Berlin (ICCE-Berlin), IEEE, 02-05 September 2018.

[5] R. Smith, "An Overview of the Tesseract OCR Engine", Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), IEEE, 23-26 September 2017

[6] M. Abdul Rahiman, M.S. Rajasree , "A Detailed Study and Analysis of OCR Research in South Indian Scripts", International Conference on Advances in Recent Technologies in Communication and Computing, IEEE, 27-28 October 2009.

[7] Mrunal G. Marne, Pravin R. Futane, Sakshi B. Kolekar, Aditya D. Lakhadive, "Identification of Optimal Optical Character Recognition (OCR) Engine for Proposed System", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE, 16-18 August 2018

[8] Muhammed Tawfiq Chowdhury, Md. Saiful Islam, Baijed Hossain Bipul, "Implementation of an Optical Character Reader (OCR) for Bengali language", International Conference on Data and Software Engineering (ICoDSE), IEEE, 25-26 November 2015.

[9] Puja Romulus, Yan Maraden, Prima Dewi Purnamasari , "An analysis of optical character recognition implementation for ancient Batak characters using K-nearest neighbors principle", International Conference on Quality in Research (QiR), IEEE, 10-13 August 2015.

[10] I. Patel, D. Patel, "Optical Character Recognition by Open-Source OCR Tool Tesseract", International Journal of Computer Applications, Volume 55-No.10, October 2017.

[11] Shiravale S. S, Sannakki S. S and Rajpurohit V. S, "Recent Advancements in Text Detection Methods from Natural Scene Images", International Journal of Engineering Research and Technology, ISSN 0974-3154, Volume 13, Number 6 (2020), pp. 1344-1352.

[12] Shiravale S. S, Sannakki S. S and Rajpurohit V. S, "Recent Advancements in Text Detection Methods from Natural Scene Images", International Journal of Engineering Research and Technology, ISSN 0974-3154, Volume 13, Number 6 (2020), pp. 1344-1352.

[13] S. S. Shiravale, R. Jayadevan and S. S. Sannakki, "Recognition of Devanagari Scene Text Using Autoencoder CNN", Electronic Letters on Computer Vision and Image Analysis, 2021, pp.55-69.

[14] S. S. Shiravale, S. S. Sannakki, R. Jayadevan, "Text Region Identification in Indian Street Scene Images Using Stroke Width Transform and Support Vector Machine" SN Computer Science, 2021.