1. INTRODUCTION

1.1 Introduction to Digital image processing

Digital Image processing performs various steps like pre-processing, extraction, restoration, acquisition, and segmentation of the image, image processing is converted analog image to digital image (with digital filters), in digital image processing extract useful information from the digital image which is already converted method to convert an image into digital form and perform some operations on it, to get an enhanced image or to extract some useful information from it. It is a type of signal dispensation in which input is an image, like a video frame or photograph and output may be image or characteristics associated with that image. Usually, the Digital Image Processing system includes treating images as two-dimensional signals while applying already set signal processing methods to them with new specifications and updated segmentation done. Digital image processing allows the use of much more complex algorithms, and hence, can offer both more sophisticated performance at simple tasks, and the implementation of methods that would be impossible by analog means. Digital image processing is the only practical technology for Classification, Feature extraction, Multi-scale signal analysis, Pattern recognition, Projection. Some techniques which are used in digital image processing include Anisotropic diffusion, Hidden Markov models, Image editing, Image restoration, Independent component analysis, Linear filtering, Neural networks, Partial differential equations, Pixilation, Principal components analysis, Self-organizing maps, and Wavelets. Digital image processing helps in accessing the information of images.

Documents are a part of our everyday life. They play an important role in helping us organize the information in various ways like categorization or classification, graphical or statistical representation, etc. A document can be defined as a unit of handwritten, printed or electronic material that offers information or serves as evidence or official record. And document image analysis helps in managing this form of information. Document image analysis is a sub-branch under Image Processing, which refers to manipulating and extracting information from a large set of data. It is a broad field of study that has been around for a long time. Research in this area has been going on for quite some time now. One of the most popular applications of this is Optical Character Recognition (OCR). There are various stages involved in Document image analysis such as pre-processing, morphological transformation, segmentation, and object recognition. Raw data must often be corrected and adjusted for errors before any data analysis or image interpretation can take place. And preprocessing is a broad term used that encompasses all techniques used to process the data collected for experimentation, usually in the form of an image. The main aim of pre-processing is to enhance the image data by suppressing unwanted distortions or enhancing certain significant characteristics in the image for subsequent processing. This includes conversion of the image from RGB to grayscale or binary form, data augmentation, (i.e., scaling, rotation, etc.), contrast or illumination correction

1.2 Objectives of Research

Developing modules for document image processing

1. Text Line segmentation for south Indian languages Documents
2. Table detection with the morphological operation
3. Linear Text Transformation for Pre-printed Documents

1.3 Introduction to Document image processing

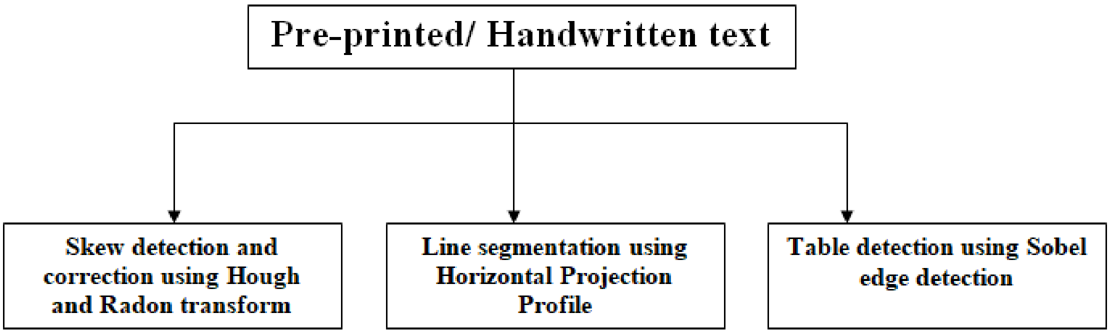
In image processing, we will be working mainly on Document Image Processing. Document image analysis refers to algorithms and techniques that are applied to images of documents to obtain a computer-readable description from pixel data. A well-known document image analysis product is the Optical Character Recognition (OCR) software that recognizes characters in a scanned document. Handwritten Text Recognition (HTR) is the ability of a computer to receive and interpret intelligible handwritten input from sources such as paper documents, photographs, touch screens, and other devices.

Handwriting plays a key role in communication. Handwritten Text Recognition (HTR) is a means by which a system perceives and interprets input which is in the form of handwritten text, i.e. scanned document. HTR is a sub-branch of Document Image Processing (DIP). It is a wide area of research that has been around for quite some time. And the field of automated HTR has achieved significant success in real-world applications. Despite the success, unconstrained HTR is still a challenge as there can be different forms of handwritten text, i.e. pen-paper handwritten text, smart device handwritten text (using a stylus or fingertip), etc. And there is a great deal of room for improvement. in terms of improved precision in the outcomes. Document analysis is one of the widely recognized and used mechanisms to scan and preserve documents

1.4 Applications

1. Optical Character Recognition improvisation
2. Improvised text document lines
3. An improvised character with skew correction and detection
4. Table Detection helps in Banking, medical sectors

1.5 Highlights of Proposed method


Pre-printed/ Handwritten text
Skew detection and correction using Hough and Radon transform
Line segmentation using Horizontal Projection Profile
Table detection using Sobel edge detection

2. LITERATURE SURVEY

2.1 Papers Reviewed

Khairum et al. [1] propose a new binarization technique to binarize an old document by addressing various factors that have led to the degradation of the documents. Their proposed method consists of 4 stages: histogram analysis, contrast enhancement, local adaptive thresholding, and artifact removal. A new way of histogram analysis is used to analyze how the pixels are deployed from an image. Con- contrast enhancement methods are used to increase the image's quality. After improving the contrast of the input image, the image is binarized using local adaptive thresholding. The next step, artifact removal, removes any noise or unwanted regions (artifacts) in an image. The proposed method performed well on binarizing single documents that were severely degraded.

Abdalla et al. [2] present a study on the various noise reduction and filtering techniques. There are various types of noises such as speckle noise, quantization noise, etc. The most commonly seen noise are impulse noise, additive noise, and multiplicative noise. There are also various methods available to filter this noise out from the images, including linear filters (adaptive filters, non-linear filters) and median filters. Since the median filter eliminates the majority of the noise in the image, the results obtained by using it improve the image's quality. But this filter increases the complexity of computation.

Xiaoyu et al. [3] present a method to rectify the document im- ages which are distorted and contain noise using patch-based Convolutional Neural Network (CNN). Their method involves first learning the flow of distortion on input image patches, then stitching the patch's effects into the rectified documents. A new network is added to fix the inconsistent illumination, which increases readability and OCR accuracy even further. Because of the less complex distortion in the smaller image patches, the proposed method's overall accuracy is greatly improved. The proposed system yields desirable outcomes for most of the documents but there might be some exceptions as they have not considered some of the conditions in a document image such as document boundary is not considered.

Zimmermann et al. [4] present an automatic segmentation system for recognizing handwritten cursive text using a hidden Markov model (HMM). An ASCII value is assigned to each word in the ASCII ground truth table, which corresponds to the bounding box of the corresponding word in the text image. And, unlike in the recognition phase, where each word is recognized, they do not recognize each word but instead, provide the Viterbi decoder with a transcript of a full line of text. This method produces excellent segmentation results, but it necessitates the use of well-trained character HMMs. The proposed method is created and tested using images of cursive handwritten English text from the IAM database. On a database with ground truth at the level of complete lines, the right word segmentation rate was 98 percentage.

Yan et al. [5] present a structure for sequence tagging and use it to segment words in a variety of languages with various writing systems and typological characteristics. The Universal Dependencies (UD) terminologies are used, with terms being elucidated as syntactic units. On the UD datasets in different languages, word segmentation is performed. The word segmentation scheme is shown as a sequence labeling model. The research uncovers a connection between language properties and segmentation accuracy in languages and writing systems. Neural network layers are used for word segmentation where each layer performs a different task. Whitespaces are used as delimiters for the segmentation of words. On the UD datasets, the proposed system has achieved state-of-the-art accuracy and outperforms previous work, especially for the most difficult languages.

Bernard et al. [6] say that the word segmentation process is done with multiple segmentation using the words-stats package, and this tool helps in prints the number of lines, word, and character in the document, and evaluation of segmentation with different levels of word token, word type, and the boundary.

Gonenc et al. [7] propose a method that works on semantic relatedness. Segmentation is based on pre-trained wordnet or simple word repetition with compared evaluation techniques. Semantic relatedness is a model which recognizes words that have some association in terms of meaning or pronunciation. For example, words such as 'Toyota' or 'Renault' are brand names, and the semantic relatedness must be able to recognize the association between the names and what they represent. Co-occurrence statistics are used to train the method, which is then learned using neural networks and linear regression.

H´elo¨ıse et al. [8] propose a new layout extraction system focused on high-level features extracted from document straight line-based segmentation. The new approach proposed is called Local Diameter Transform. This transform can be applied to either foreground or background pixels. To apply the transform various steps have to be followed such as pre-processing, document classification, and recognizing of borders present. The pre-processing step includes binarization of the image, the documents are classified based on graphical or non-graphical regions (textual and non-textual regions). After categorizing the documents, analyzed with table format data, some of the tables have a border or borderless managing layout of the document with a long straight line and short line algorithm with aggregating both algorithms to get the final result.

Kartik et al. [9] present a modified framework for CNN-RNN hybrid architecture for effective network initialization using synthesized data for pretraining, image normalization and slant correction, and data transformation and distortion. The datasets considered are IAM, RIMES, and GW. The hybrid network consists of multiple layers where each layer performs a specific task. Pre-processing takes place on both word-level and line-level recognition using slant and slope normalization. Synthetic data is used for the training phase. Various augmentation schemes are used in the process of training the system.

Basilios et al. [10] present a novel method for automatically detecting tables in images of documents, which includes horizontal and vertical line detection. The input images are pre-processed by converting them into grayscale images and finally, thresholding is done. Morphological operations are used to detect vertical and horizontal lines. Images and text in the image are omitted to increase the accuracy of the final result. Finally, the table is detected by detecting all the intersections and the table is reconstructed using the detected horizontal and vertical line.

Saman et al. [11] present an improved version of the existing deep learning methods used for table detection. Coloration and Euclidean distance measure is used for segregating the text regions and non-text regions for table detection. After preprocessing, the R-CNN model of deep learning is used to detect the tables in a document image. The proposed work is performed on the publicly available dataset UNLV.

Bipin et al. [12] propose an offline Handwritten Malayalam character recognizer using OCR. Their proposed method mainly consists of 2 stages - pre-processing and segmentation. Pre-processing is done using Otsu, Sauvola, and Hybrid methods. Word level and character level segmentation is done in the segmentation phase, i.e. the document image is split into two parts - all the words in the image are segmented and stored; all the characters presented are segmented and stored. The accuracy comparison of binarization is done using PSNR.

Bipin et al.[13]propose a comparative study for binarization of aged his- historical documents of agreements (75-100 years old documents) using local and global thresholding. Otsu, Niblack, sauvola, wolf methods are used for the comparative study along with morphological operations such as dilation and erosion for a clearer document image. Kannada and Malayalam datasets are used for the experimentation. SNR and MES methods are used for the evaluation of the results obtained.

N. Shobha Rani et al.[14]propose a method for pre-processing to detect and remove horizontal and vertical lines in pre-printed documents. Their proposed approach is divided into two stages - the first phase consists of image enhancement and line detection (Laplacian operator is used for edge detection); the second phase consists of text stroke crossings on lines detection using rectangular structuring element. 60 Telugu script document images are used as a dataset for experimentation.

Sreelakshmi U.K et al. [15]propose a technique for variable regions detection from aged documents using object detection and nearest neighbor classifier. For object detection, connected component analysis is used, and Histogram of Gradient (HoG) features are used to differentiate between printed text regions and variable regions in an image. Experimentation is based on datasets Tobacco800 Complex Document Image Database. Based on the observation made in the related work and after studying the already existing techniques, we propose a system that uses touching or overlapping regions as factors for the line segmentation, morphological operations to perform table detection. This system is applicable in various fields – digitizing handwritten and scanned documents, medical NLP, banking, and insurance sector, etc

C.S. Vijayashree et al. [16] propose a model for improved OCR readability by using text images as input which is in the form of a wave. The input is subjected to character segmentation followed by tilt correction and correction of character alignment. The proposed work efficiently transforms wave-form-text into a linear-form text. The performance of the proposed work is considered good with the OCR readability of 98% after transformation.

Avinash et al. [17] present an online skew detection and correction algorithm. Brenham line drawing algorithm is used to find the angle for skew correction and once the angle is found, the tilt is corrected using hough transform.

Ibrahim et al. [18] present two algorithms for skew detection and correction on document images. These algorithms make use of the horizontal RLSA image of a skewed document. An average of the selected area in the document's black connected component is used as the base angle for tilt correction by rotating the whole document in the opposite direction of the angle obtained.

Yang et al. [19] present a new skew detection method based on straight-line fitting, which used Eigen points to detect the angle of the tilt in the document. An Eigen-point is the bottom center of the bounding box of a connected component. The algorithm works using the selection of a sub-region and the objects required for analysis. An analysis is conducted based on the selected region between the relations between neighboring Eigen-points in every text line within a suitable sub-region.

Chandan et al. [20] present a paper on the time complexity analysis on the various stages of the skew correction process. The skew correction process includes three parts – preprocessing using a simplified Block Adjacency Graph, slant detection using hough transform with four variants of the same and finally skew correction using forward rotation, inverse rotation, and Bresenham's line algorithm.

Deepak et al. [21] present a new method of Hough transform to reduce the time consumed for processing without reducing the accuracy of the method. The HT space continuum is divided, i.e., the angle of skew between degrees 0 and 45, into one-tenths, which yields the portion in which the resultant skew lies.

Jian et al. [22] present Radon transform-based two algorithms for word slant and skew correction. The Radon transform is used for skew correction by maximizing the global measure that is defined by the Radon transform and to estimate the long strokes for the slant correction, and the average angle of these long strokes is utilized to calculate the word slant. The tests were carried out on the IMDS cursive word database.

Bogdan et al. [23] present a skew detection system using Radon transform. The proposed method summarizes the image as input filters out the larger figures in the image and keeps with a larger width which is useful for skew detection. Later the image is tested for various angles and the most suitable angle is recorded and retained.

Daniel et al. [24] present a paper on the analysis of the problems faced during skew detection and correction by implementing various algorithms such as Hough and Radon transform for skew detection and correction along with SLIDE (Subspace-based Line Detection) for line detection.

Lipi et al. [25] propose a skew detection and correction method on handwritten and pre-printed Gujarati documents. The proposed method works on the Linear Regression method for detecting the angle of rotation for skew correction. An accuracy of 59.63% and 45.58% is obtained for printed and handwritten documents respectively. The proposed method fails to provide results for small variations in the angle of rotation while giving good results in the case of major angles of rotation.

E. Del Ninno et al. [26] present a projection-based non-linear de-skewing algorithm. The de-skewing works in 3 stages – select a section of the document; estimate the angle of skew and finally correct the skew. The proposed work is based on a projection profile for lines consisting of spaces and lines consisting of text.

Salem et al. [27] present a skew detection and skew correction for Quran page images based on the Hough transform. The skew angulation is calculated using the Hough transform lines detection technique. The proposed method works in the following stages – pre-processing (consists of input image conversion into grayscale, then into binary and later foreground image detection); line detection using Hough transform and skew angle detection; followed by image skew correction using the angle detected. It works best between 0 and 20 degrees of skew angle for skew detection and correction.

Amani et al. [28] present a method for skew correction of Arabic handwritten scripts. The proposed method has shown high accuracy for skew detection and has given better results than state-of-the-art techniques.

Omar et al. [29] present a novel method for skew angle detection and correction. The proposed method works in the following stages – image binarization, morphological skeleton extraction, Probabilistic Hough Transform line detection, skew angle estimation, and finally skew correction. The model was tested on 3 different types of datasets – document type dataset (books, maps, etc.), different linguistic documents dataset, and finally documents with different layouts and alignments. The proposed method provides efficient and accurate results and a large number of variables; this mentation can only show the global image slope of the document.

B. Eswara Reddy [30] presents a novel algorithm for skew detection on Telugu documents. The proposed method is based on the Principle axis Farthest Pairs Quadrilateral (PFPQ) on any document that contains textual, tabular, or pictorial regions. This method applies to any document and has no angle restrictions.

Bezmaternykh et al. [31] present a skew detection algorithm based on Fast Hough Transform analysis. This approach uses FHT calculation for both horizontal and vertical lines and does not require an initial binarization step for the input image, and the reduced computational cost of the FTH calculation broadens its usefulness. DISEC’13 dataset is used for experimentation.

2.2 Motivation

HTR is a very fascinating area to work in as it has endless opportunities for development and many areas with room for improvement. Research into this area is imperative as the world is fast approaching an era of complete digitization of data including handwritten text documents in different languages - western and regional. Different languages have different forms and also there are numerous ways a person can write the language, i.e. different forms of handwriting. Due to this form of focus/challenge, we have chosen this area of research.

2.3 Conclusion of the Survey

Various methods for image processing have been introduced in the past years. These techniques play an important role in manipulation and extraction of information from the images being processed. The survey conducted highlights the different techniques for document image analysis which includes methods for pre-processing, segmentation, table detection, etc.

3. Automated Text line Segmentation and Table detection for Pre-Printed Document Image Analysis Systems

3.1 Introduction

Handwriting plays a key role in communication. Handwritten Text Recognition (HTR) is a means by which a system perceives and interprets input which is in the form of handwritten text, i.e. scanned document. HTR is a sub-branch of Document Image Processing (DIP). It is a wide area of research that has been around for quite some time**.** And the field of automated HTR has achieved significant success in real-world applications. Despite the success, unconstrained HTR is still a challenge as there can be different forms of handwritten text, i.e. pen-paper handwritten text, smart device handwritten text (using a stylus or fingertip), etc. And there is a lot of scope for improvement in terms of improved precision in the outcomes.

Document analysis is one of the widely recognized and used mechanisms to scan and preserve documents [14]. It helps in organizing and recognizing textual and non-textual regions. Researchers have been exploring this domain and its branches for quite some time now. A well-known document image analysis product is Optical Character Recognition (OCR) software that recognizes characters in a scanned document. There are various ways in which image processing or document analysis can be performed. It includes text enhancement, baseline detection, line segmentation, word segmentation, table detection, etc.

In this work, we investigate the contribution of models for line segmentation and table detection. Line segmentation and table detection play an essential role in document analysis. Line segmentation refers to dividing an input image based on the textual lines present in the image. It is usually performed on document images - printed or handwritten documents. Table detection refers to the detection of a structure which is a collection of horizontal and vertical lines along with intersections. This form of detection can be applied in different fields where document is a regular part of work, especially in the banking and insurance sector.

Line segmentation and table detection is a challenging area of research as there are different forms of documents – handwritten and printed. And all these documents have different layouts. In the case of handwritten documents, the text is not written consistently. For printed document images, the quality and the skew of the image depends on how the image has been scanned, i.e., the pages scanned need not be straight which makes it difficult for the system to recognize characters and structures.

Therefore in this paper, we aim at devising algorithmic models suitable for overlapping line segmentation in handwritten documents and table detection in pre-printed documents.

3.2 Working of the proposed method

In the proposed method 2 different models are devised. The first model uses projection profiles to segment overlapping lines from handwritten document images. In the latter, a model for table detection in pre-printed document images is proposed.

3.2.1 Architecture of Text line segmentation

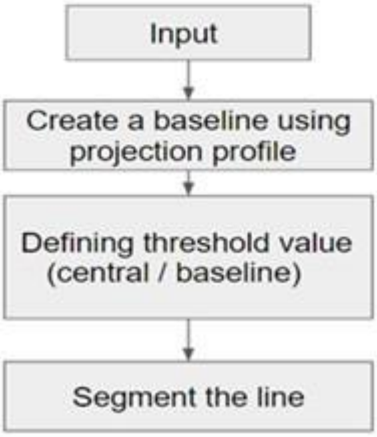

Input
Create a baseline using projection profile
Defining threshold value (central / baseline)
Segment the line

Fig 3.2.1.1 Architecture of Text Line Segmentation

Initially, the handwritten document image is assumed as input to the proposed model. The inputs employed for overlap-ping line segmentation are collected by synthetically obtained handwritten documents. Assumption of synthetic data creation includes the presence of only overlapping lines in south Indian scripts such as Kannada, Telugu, and Malayalam. Datasets are collected from more than 100 writers from various regions of south India. The overall dataset comprises 235 handwritten document images. The datasets consist of 25-30percentage of each script type. Input document images assumed for line segmentation are subjected to grayscale image conversion followed by binarization. The projection profile is computed after an image has been binarized. A projection profile is a one-dimensional vector that can be used to represent the distribution of foreground intensities (textual information) for a specific row or column. The row-wise distribution of foreground intensity refers to the horizontal projection profile, whereas the column-wise distribution refers to the vertical projection profile. In the approach proposed, the baseline is created using the one-dimensional vector computed from the projection profile.

3.2.2 Architecture of Table detection

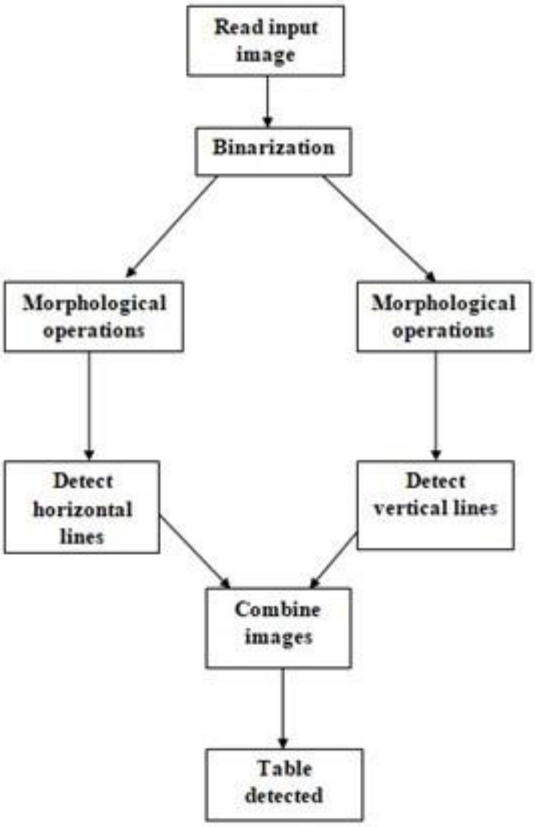
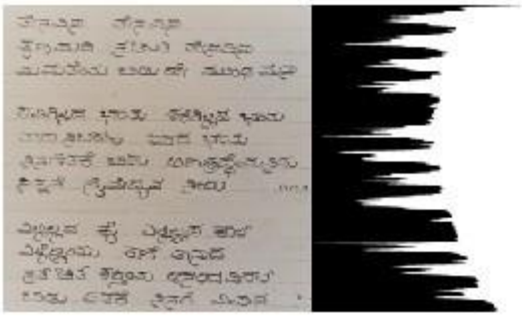


Fig.3.2.2.1 Architecture of Table Detection

Pre-printed document images are assumed as input and subject to thresholding. The outcome of the thresholding is a binarized image that is submitted for morphological processing via erosion followed by dilation. shows the output of morphological processing for the input samples. Applying morphological operation may result in the enhancement of gradient details including text and graphical elements. This also helps in the accurate detection of tables from pre-printed document images.



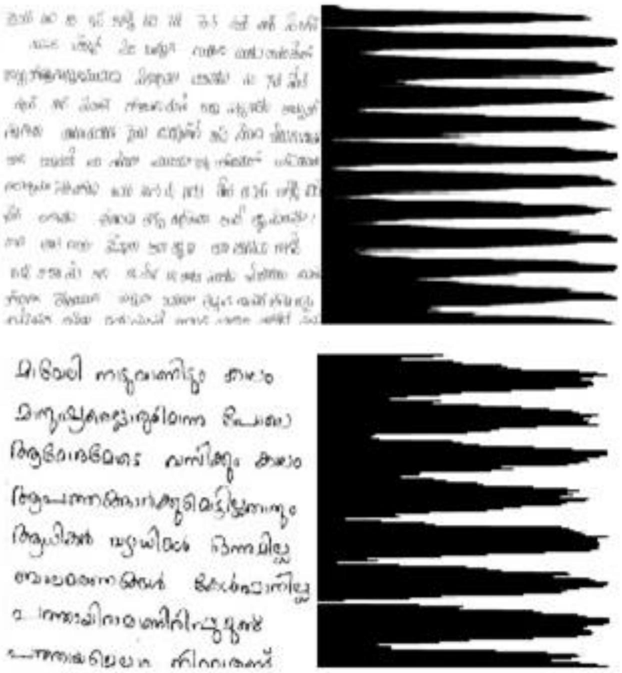


Fig. 3.2.1 Handwritten document sample - horizontal projection profile

Once the baseline is determined from the projection profile, it is employed for the extraction of overlapping lines. If V1, V2,..., Vk represents the value between the peaks P1 and P2 in an image then, the valley Vi obtained from a minimum of V1, V2,..., Vk is considered as the baseline. Based on the baseline ‘b’, a threshold of +/- t is assumed to have from top or bottom from a position p of baseline b. Each point on the baseline is a pixel present in the corresponding position of baseline b. All the pixels in that particular baseline position are interpreted to check the presence of any overlapping characters that are occurring in between 2 lines. The presence of a particular pixel belonging to an overlapping character will fall in the range of p+ t and p-t. The boundary position of a pixel is assumed to be present with a consecutive occurrence of k pixels with background intensity (non-textual pixels).

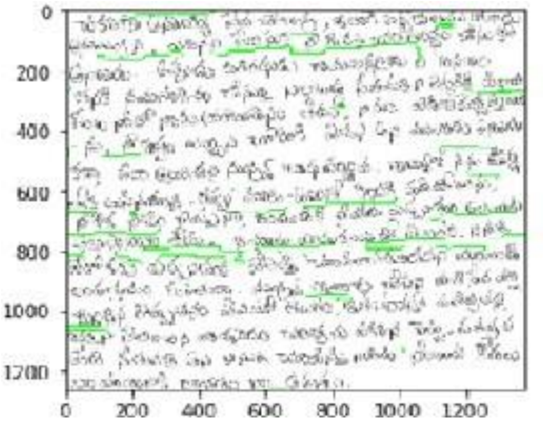
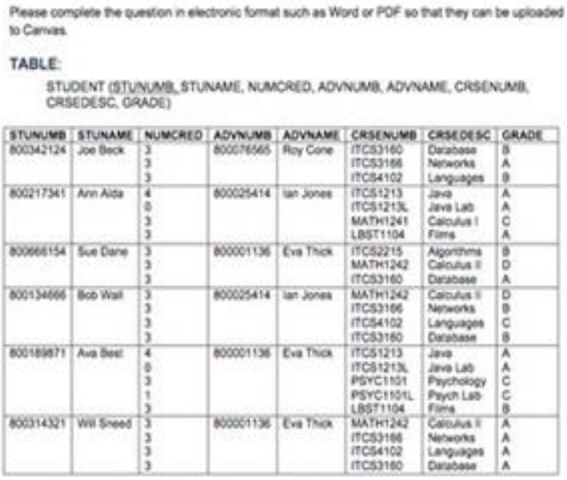


Fig. 3.2.2: Sample images with line boundaries detected - overlapping lines

Thus, the proposed line segmentation method can extract the overlapping lines from handwritten document images. In the subsequent section, the technique proposed for table detection is illustrated.



Please complete the question in electronic format such as Word or PDF so that they can be uploaded to Canvas.

TABLE:

STUDENT (STUNUMB, STUNAME, NUMCRED, ADVNUMB, ADVNAME, CRSENUMB, CRSEDESC, GRADE)

STUNUMB	STUNAME	NUMCRED	ADVNUMB	ADVNAME	CRSENUMB	CRSEDESC	GRADE
800342124	Joe Beck	3 3 3	800076565	Roy Cone	ITCS3160 ITCS3166 ITCS4102	Database Networks Languages	B A B
800217341	Ann Alda	4 0 3 3	800025414	Ian Jones	ITCS1213 ITCS1213L MATH1241 LBST1104	Java Java Lab Calculus I Films	A A C A
800666154	Sue Dane	3 3 3	800001136	Eva Thick	ITCS2215 MATH1242 ITCS3160	Algorithms Calculus II Database	B D A
800134666	Bob Wall	3 3 3 3	800025414	Ian Jones	MATH1242 ITCS3166 ITCS4102 ITCS3160	Calculus II Networks Languages Database	D B C B
800189871	Ava Best	4 0 3 1 3	800001136	Eva Thick	ITCS1213 ITCS1213L PSYC1101 PSYC1101L LBST1104	Java Java Lab Psychology Psych Lab Films	A A C C B
800314321	Will Sneed	3 3 3 3	800001136	Eva Thick	MATH1242 ITCS3166 ITCS4102 ITCS3160	Calculus II Networks Languages Database	A A A A

Fig.3.2.2.1 Input image for table detection

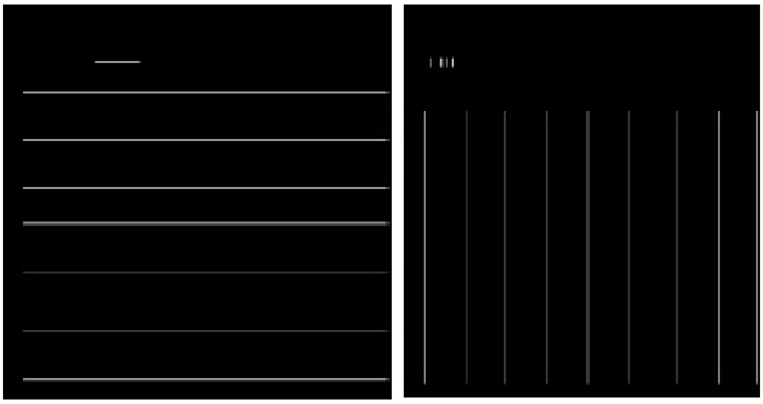


Fig.3.2.2.2 after the morphological operation

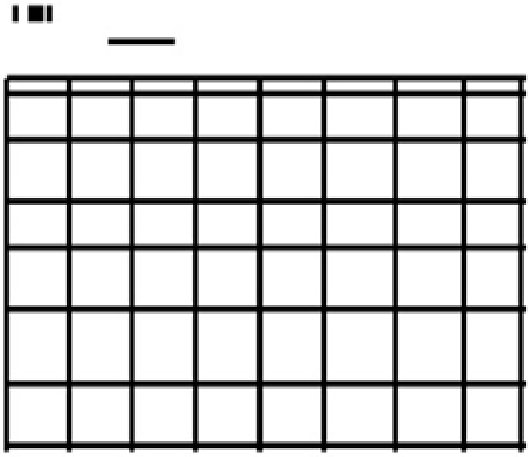
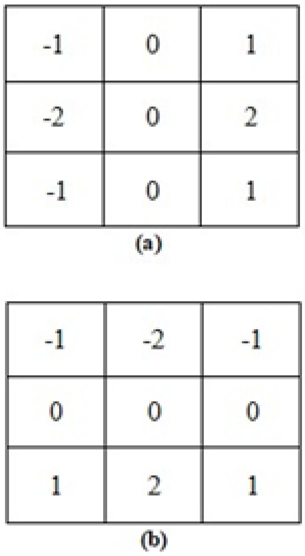


Fig.3.2.2.3 outcome of table detection

A. *EDGE DETECTION*

Detection of intensity discontinuities in the process of identifying the direction and magnitude in which the edges are defined in the images. In the system proposed, the Sobel edge detection method is used for table detection. Sobel edge detection is crucial in the detection of the horizontal and vertical oriented edges. The difference in one pixel to another pixel intensity values in the horizontal and vertical direction is computed using Sobel operator kernels with the change of specific weights.



-1	0	1
-2	0	2
-1	0	1

(a)

-1	-2	-1
0	0	0
1	2	1

(b)

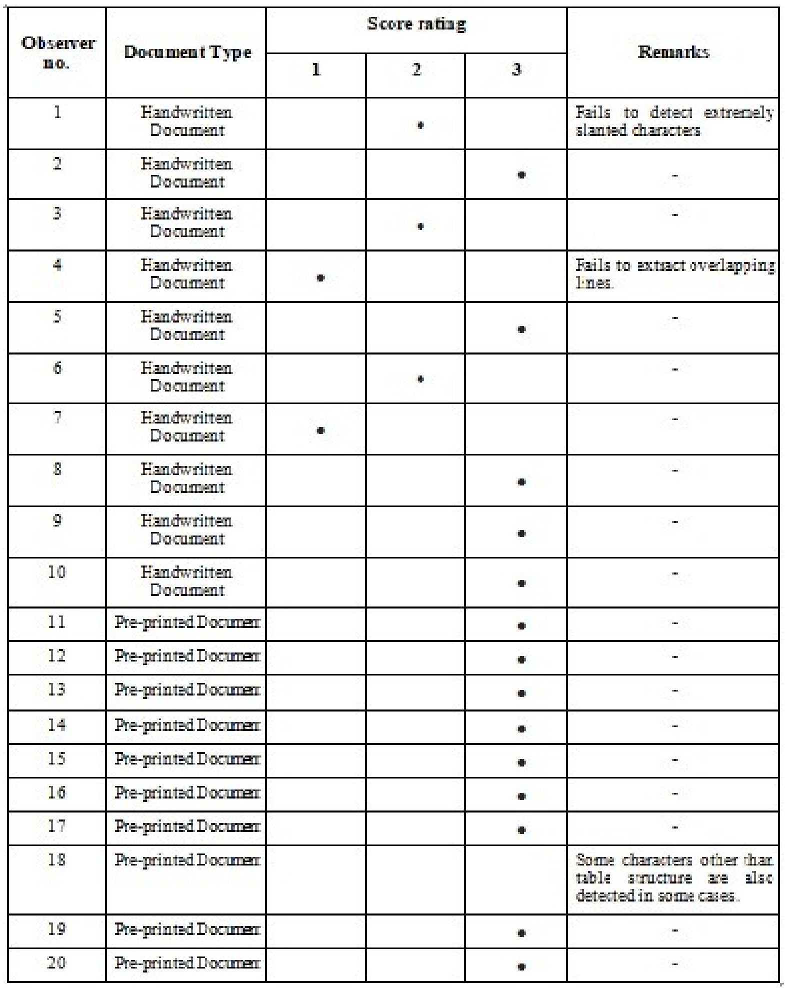
Fig. 3.2.2.4 Sobel edge operator for horizontal and vertical edge detection

In the proposed method, a kernel of size 128 x 128 is used for the detection of horizontal and vertical edges. Input to the Sobel edge detector is assumed as a binarized image. So, once the edges are detected, the outcome of the table-detected image is obtained.

3.3 Experimental analysis

In the proposed work, 355 images are assumed as samples for experimentation, consisting of both handwritten document (235) and pre-printed document (120) images. Evaluation of the proposed work is conducted through non-reference data validation methods. The outcomes of line segmentation and table detection techniques are validated with the help of 20 observers. Each observer with data evaluation forms can rate the outcomes obtained in terms of ratings from 1 to 3. Higher ratings indicated good results, 2 indicated satisfactorily, and 1 for poor results. The outcomes of validation provided by the observers are shown in Table 1.

From Table 1, it is evident that 70 percent of document images are validated successfully with a score of 3. This implies the efficiency of the proposed method, however, the ‘line segmentation technique proposed fails to extract the touching lines and also lines with highly variant and slanted characters.



Observer no.	Document Type	Score rating			Remarks
		1	2	3	
1	Handwritten Document		•		Fails to detect extremely slanted characters
2	Handwritten Document			•	-
3	Handwritten Document		•		-
4	Handwritten Document	•			Fails to extract overlapping lines.
5	Handwritten Document			•	-
6	Handwritten Document		•		-
7	Handwritten Document	•			-
8	Handwritten Document			•	-
9	Handwritten Document			•	-
10	Handwritten Document			•	-
11	Pre-printed Documen			•	-
12	Pre-printed Documen			•	-
13	Pre-printed Documen			•	-
14	Pre-printed Documen			•	-
15	Pre-printed Documen			•	-
16	Pre-printed Documen			•	-
17	Pre-printed Documen			•	-
18	Pre-printed Documen				Some characters other than table structure are also detected in some cases.
19	Pre-printed Documen			•	-
20	Pre-printed Documen			•	-

Table:3.3.1 rating of outcome

3.4 Conclusion

In the proposed model, line segmentation from the south Indian script is achieved using horizontal projection and has an accuracy of 85-90 percent. In addition to line segmentation, active contour regions have also been to remove the touching lines as well as detection of tables from documents using morphological operations. The main drawback of the work is in the case of skewed documents. Line segmentation is a challenging thing so our proposed work has failed in the case of skew correction as the kernel value varies which in turn affects the detection of tables from the documents. In the future, we can make use of a skew correction algorithm before line segmentation for such documents. To overcome the problem of varying kernels, pixel-by-pixel mean value calculation is used to fix the kernel value.

4. Proposed method for Skew Detection and Correction

4.1 Working of proposed method

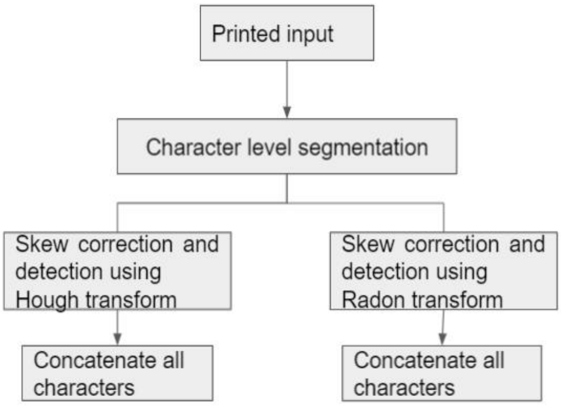

Printed input
Character level segmentation
Skew correction and detection using Hough transform
Skew correction and detection using Radon transform
Concatenate all characters
Concatenate all characters

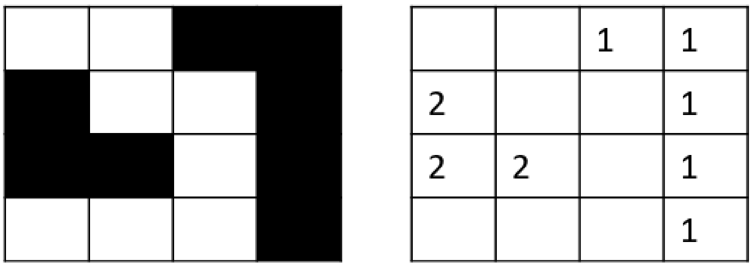
Fig . 4.1.1 Architecture for Skew detection and correction

Text images with skew are considered input, i.e., text in waves, text in the diagonal direction, etc. The input undergoes character level segmentation using connected components. After the character segmentation, the segmented characters undergo grayscale conversion, and the Canny edge detection method is used for detecting the edges of the character

4.2 Experimental analysis

A. Connected components analysis

The labeling method will help gather a connected pixel based on 4-way or 8-way. Most of the connectivity uses 8-way pixel accessing, and to connect the next connecting pixel, it’s based on heuristic analysis. For character segmentation, the connected component will give good results, because the binary image array will have an intensity map so that they may be connected. Fig. 3 shows an example of the mapping of a pixel in the connected component.



		1	1
2			1
2	2		1
			1

Fig. 4.3.1 Example Mapping of a pixel in the connected component

B. Canny edge detection

This algorithm will help in identifying the edges of the character, canny is a multiple-step algorithm considering a Gaussian filter for noise removal, Sobel kernel to the smoothening image then suppress the background pixels these above steps will give a thin-edged image, hysteresis is helping in indenting the maximum value of thresholded edge.

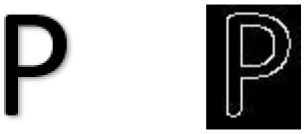


Fig.4.3.2 Result image of edge detection

C. Skew correction and detection

In Optical Character Recognition(OCR), one of the major problems is segmentation, before that we will consider slant/tilt angle in text documents or regions of the document, skew is that it will not be in a positive angle, the positive angle may be parallel or intersected. As per the baseline of text or pixel, it will be calculated at the angle and will be corrected using rotation.

D. Hough Transform

A method will detect a line based on an edge map and the baseline of an edge will be considered as a line, these lines are plotted on an empty image array and will detect the angle between those lines.

$$A \cos \theta + B \sin \theta = \text{angle}$$

```
Algorithm
    1. Input
    2. Binary image
    3. Loop in every pixel fun(x,y):
                If $0 <= \theta$ AND $\theta < 180$:
                        $A \cos \theta + B \sin \theta = angle$
                add(angle, $\theta + 1$)
        End loop
```

E. Radon Transform

A method is helping in identifying the text baseline angle through the original domain of the image. This algorithm considers the collection points as a referral in image

$$p(\emptyset, sa) = \int_l (a,b)dl$$

All points of the image on this line satisfy the equation

$A*\sin(\emptyset) - B*\cos(\emptyset) = sa$

4.3 Conclusion

Skew correction is an essential part of document image analysis as a non-zero skew document or dataset may cause problems in the further stages of document analysis. Hence, to overcome this problem, we propose a skew correction model using Hough transform and Radon transform based on text images that have been separated using Character segmentation. The proposed model gives an accuracy of

88-95 percent of accuracy. The segmented characters undergo grayscale conversion before skew correction. The main drawback of the work is it is mainly tested on text lines only and not the image of a whole document.

5. CONCLUSIONS

5.1 Summary

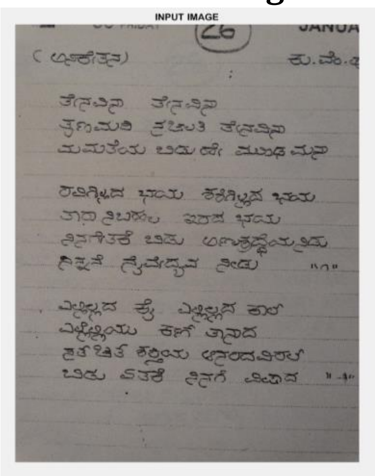
The proposed work consists of several methods and techniques used in Document Image Analysis such as skew correction techniques, segmentation, and edge detection techniques. The skew correction technique which falls under preprocessing is used for slant correction in-text images using the segmented characters from the text images. The slant correction is done using two methods Hough transforms and Radon transforms along with edge detection techniques for character recognition after segmentation. The next technique that has been experimented on is Line segmentation. This is done on South Indian Handwritten script images using Horizontal Projection Profile for baseline detection and segmentation. Finally, the last module of this experimentation is table detection. This is performed on pre-printed document images. The table detection is carried out in the following manner - horizontal and vertical lines in the image are detected using separately after performing morphological operations (dilation and erosion) using Sobel edge detection technique, and later on, the vertical edge detected image and horizontal edge detected image are merged to give the final output of table detected from the input image. The above-mentioned techniques are used for experimentation to overcome certain challenges faced in each module. Skew correction is an essential part of document image analysis as a non-zero skew document or dataset may cause problems in the further stages of document analysis, line segmentation poses a challenge in extracting the overlapping lines from handwritten document images.

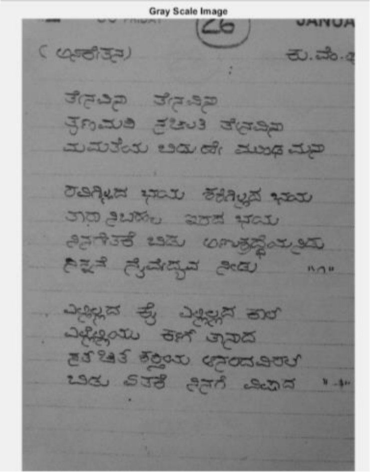
5.2 Future directions

These methods used are minor improvisations of the existing techniques with slightly increased accuracy and clear results. In the future, we could work on coming up with a new algorithm based on the existing methods and techniques to increase the accuracy to the maximum with as little inefficiency as possible.

ANNEXURE

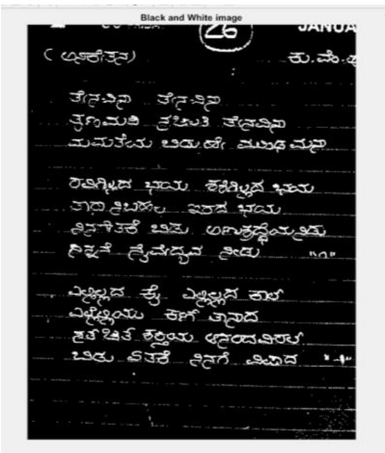
1. The outcome of Text line segmentation

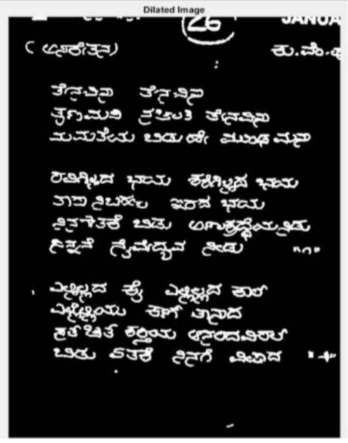




Input Image

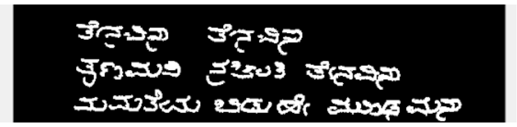
Gray Scale image



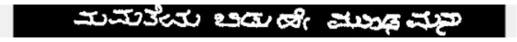


Binarized

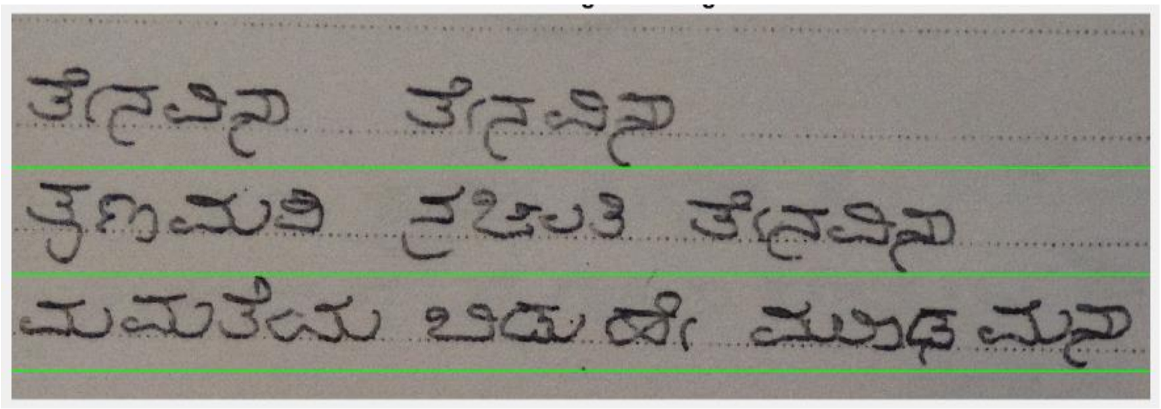
Noise Removed



Selected Area for Line Segmentation

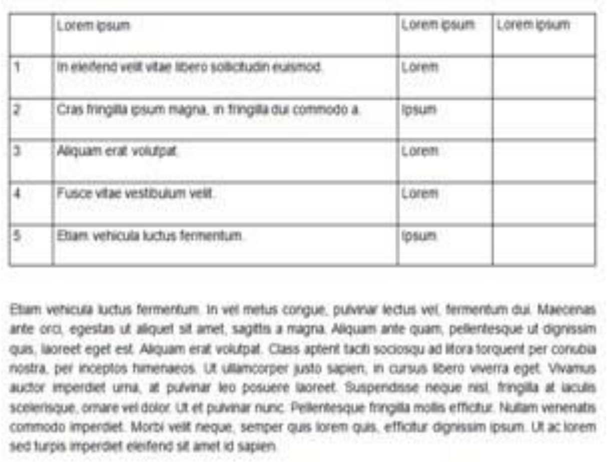


The segmented text line image



The text line is drawn

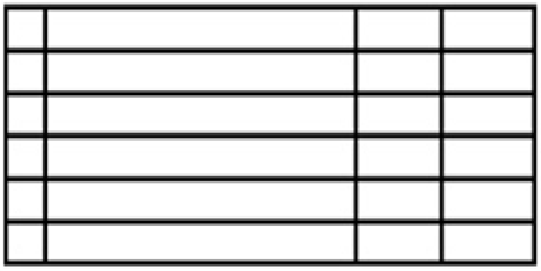
2. The outcome for Table Detection



	Lorem ipsum	Lorem ipsum	Lorem ipsum
1	In eleifend velit vitae libero sollicitudin euismod.	Lorem	
2	Cras fringilla ipsum magna, in fringilla dui commodo a.	Ipsum	
3	Aliquam erat volutpat.	Lorem	
4	Fusce vitae vestibulum velit.	Lorem	
5	Etiam vehicula luctus fermentum.	Ipsum	

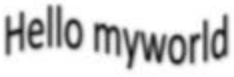
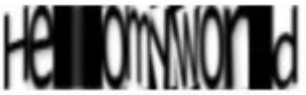
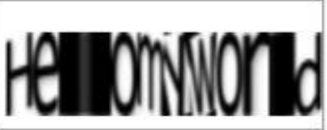
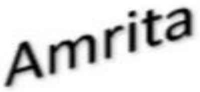
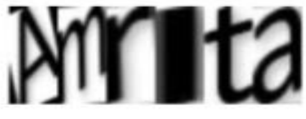
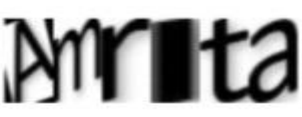
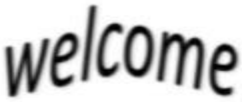
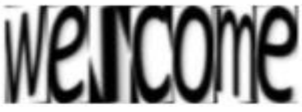
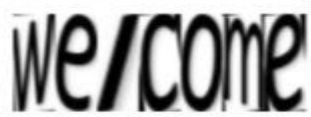
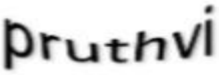
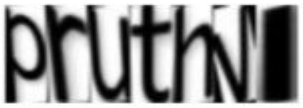
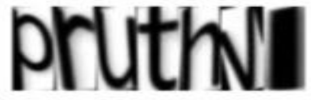
Etiam vehicula luctus fermentum. In vel metus congue, pulvinar lectus vel, fermentum dui. Maecenas ante orci, egestas ut aliquet sit amet, sagittis a magna. Aliquam ante quam, pellentesque ut dignissim quis, laoreet eget est. Aliquam erat volutpat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Ut ullamcorper justo sapien, in cursus libero viverra eget. Vivamus auctor imperdiet urna, at pulvinar leo posuere laoreet. Suspendisse neque nisl, fringilla at iaculis scelerisque, ornare vel dolor. Ut et pulvinar nunc. Pellentesque fringilla mollis efficitur. Nullam venenatis commodo imperdiet. Morbi velit neque, semper quis lorem quis, efficitur dignissim ipsum. Ut ac lorem sed turpis imperdiet eleifend sit amet id sapien.

input image

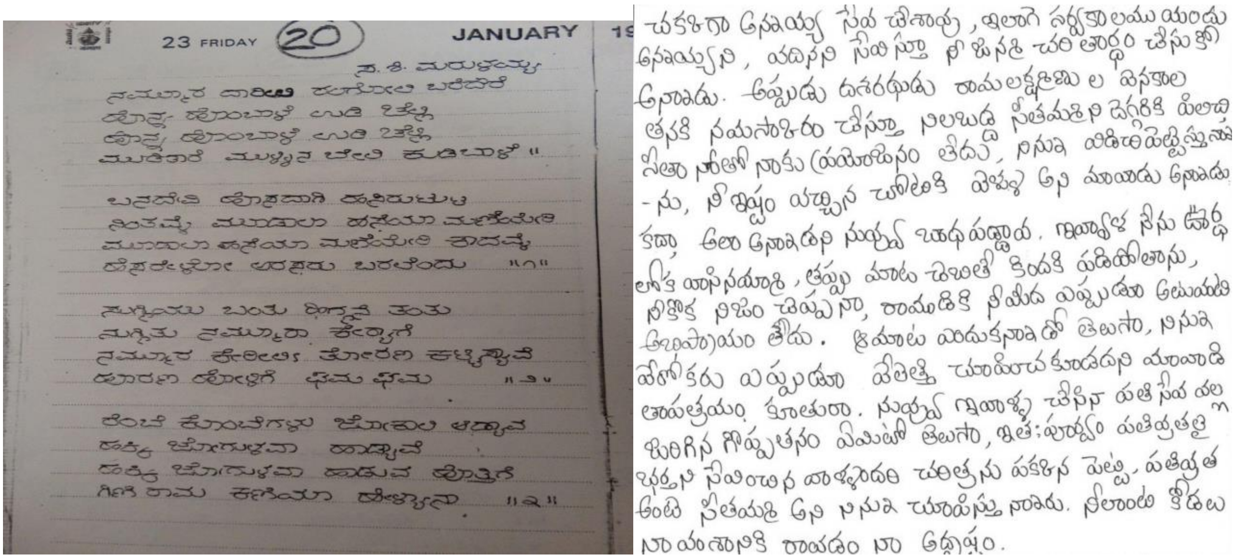


The output of the detected table

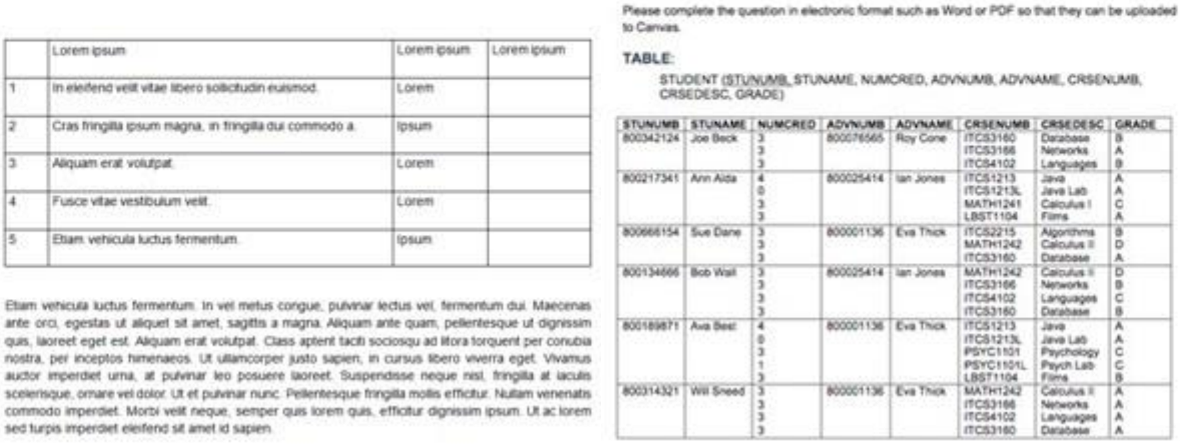
3. The outcome of Skew Detection and Correction

Input image	Hough transform	Radon transform
		
		
		
		

Dataset Samples



Text line Segmentation



	Lorem ipsum	Lorem ipsum	Lorem ipsum
1	In eleifend velit vitae libero sollicitudin euismod.	Lorem	
2	Cras fringilla ipsum magna, in fringilla dui commodo a.	Ipsum	
3	Aliquam erat volutpat.	Lorem	
4	Fusce vitae vestibulum velit.	Lorem	
5	Etiam vehicula luctus fermentum.	Ipsum	

Etiam vehicula luctus fermentum. In vel metus congue, pulvinar lectus vel, fermentum dui. Maecenas ante orci, egestas ut aliquet sit amet, sagittis a magna. Aliquam ante quam, pellentesque ut dignissim quis, laoreet eget est. Aliquam erat volutpat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Ut ullamcorper justo sapien, in cursus libero viverra eget. Vivamus auctor imperdiet urna, at pulvinar leo posuere laoreet. Suspendisse neque nisl, fringilla at iaculis scelerisque, ornare vel dolor. Ut et pulvinar nunc. Pellentesque fringilla mollis efficitur. Nullam venenatis commodo imperdiet. Morbi velit neque, semper quis lorem quis, efficitur dignissim ipsum. Ut ac lorem sed turpis imperdiet eleifend sit amet id sapien.

Please complete the question in electronic format such as Word or PDF so that they can be uploaded to Canvas.

TABLE:

STUDENT (STUNUMB, STUNAME, NUMCRED, ADVNUMB, ADVNAME, CRSENUMB, CRSEDESC, GRADE)

STUNUMB	STUNAME	NUMCRED	ADVNUMB	ADVNAME	CRSENUMB	CRSEDESC	GRADE
800342124	Joe Beck	3 3 3	800076565	Roy Cone	ITCS3160 ITCS3166 ITCS4102	Database Networks Languages	B A B
800217341	Ann Alda	4 0 3 3	800025414	Ian Jones	ITCS1213 ITCS1213L MATH1241 LBST1104	Java Java Lab Calculus I Films	A A C A
800666154	Sue Dane	3 3 3	800001136	Eva Thick	ITCS2215 MATH1242 ITCS3160	Algorithms Calculus II Database	B D A
800134666	Bob Wall	3 3 3 3	800025414	Ian Jones	MATH1242 ITCS3166 ITCS4102 ITCS3160	Calculus II Networks Languages Database	D B C B
800189871	Ava Best	4 0 3 1 3	800001136	Eva Thick	ITCS1213 ITCS1213L PSYC1101 PSYC1101L LBST1104	Java Java Lab Psychology Psych Lab Films	A A C C B
800314321	Will Sneed	3 3 3 3	800001136	Eva Thick	MATH1242 ITCS3166 ITCS4102 ITCS3160	Calculus II Networks Languages Database	A A A A

Table Detection

Amrita
pruthvi
welcome
Hello myworld

Skew Detection and Correction

Tool Used

1. MATLAB 2020R
2. Spyder 5.0; Python 3.7

Bibliography

[1] Saddami, K., Munadi, K., Away, Y., Arnia, F., ”Effective and fast binarization method for combined degradation on ancient documents”,Heliyon, 5(10), e02613, 2019

[2] Hambal, A. M., Pei, Z., Ishabailu, F. L., ”Image noise reduction and filtering techniques”, International Journal of Science and Research (IJSR), 6(3), 2033-2038, 2017

[3] Li, Xiaoyu, Bo Zhang, Jing Liao, Pedro V. Sander. ”Document rectification and illumination correction using a patch-based CNN.” ACM Transactions on Graphics (TOG) 38, no. 6 (2019): 1-11..

[4] Zimmermann, Matthias, Horst Bunke. ”Automatic segmentation of the IAM off-line database for handwritten English text.” In Object recognition supported by user interaction for service robots, vol. 4, pp. 35-39. IEEE, 2002.

[5] Shao, Yan, Christian Hardmeier, Joakim Nivre. ”Universal word segmentation: Implementation and interpretation.” Transactions of the Association for Computational Linguistics 6 (2018): 421-435.

[6] Bernard, Mathieu, Roland Thiolliere, Amanda Saksida, Georgia R. Loukatou, Elin Larsen, Mark Johnson, Laia Fibla . ”WordSeg: Standardizing unsupervised word form segmentation from the text.” Behavior research methods 52, no. 1 (2020): 264-278.

[7] Ercan, Gonenc, Ilyas Cicekli. ”Topic segmentation using word-level semantic relatedness functions.” Journal of Information Science 42, no. 5 (2016): 597-608.

[8] Alh e´ritie`re, He´lo¨ıse, Florence Cloppet, Camille Kurtz, Jean-Marc Ogier, Nicole Vincent. ”A document straight line based segmentation for complex layout extraction.” In 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 1126-1131. IEEE, 2017.

[9] Dutta, Kartik, Praveen Krishnan, Minesh Mathew, C. V. Jawahar. ”Improving cnn-rnn hybrid networks for handwriting recognition.” In 2018 16th international conference on frontiers in handwriting recognition (ICFHR), pp. 80-85. IEEE, 2018.

[10]Gatos, Basilios, Dimitrios Danatsas, Ioannis Pratikakis, Stavros J. Perantonis. ”Automatic table detection in document images.” In International Conference on Pattern Recognition and Image Analysis, pp. 609-618. Springer, Berlin, Heidelberg, 2005.

[11] Nair, BJ Bipin, C. Anil Krishnan, Joshua Johns, BV Sadris Jain, B. Sarath. ”An Enhanced Approach for Binarizing and Segmenting Degraded Ayurvedic Medical Prescription.” International Journal of Pharmaceutical Research 12, no. 3 (2020): 783-790.

[12]B.J. Bipin Nair, K.P. Nihar, C.K. Adarsh, ”A Comparative Binarization Approach for Degraded Agreement Document Image from Various Pharmacies”, International Journal of Pharmaceutical Research, 2020, DOI:10.31838/ijpr/2020.12.02.0124

[13]N. Shobha Rani and T. Vasudev, ”An Efficient Technique for Detection and Removal of Lines with Text Stroke Crossings in Document Images”, Proceedings of International Conference on Cognition and Recognition, 2018, DOI: 10.1007/978-981-10-5146-39

[14]U. K. Sreelakshmi, V. G. Akash and N. S. Rani, ”Detection of variable regions in complex document images,” 2017 International Conference on Communication and Signal Processing (ICCSP), 2017, pp. 0807-0811, DOI: 10.1109/ICCSP.2017.8286476.

[15]M. Javed, P. Nagabhushan and B. B. Chaudhuri, ”Extraction of line- word-character segments directly from run-length compressed printed text-documents,” 2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013, pp. 1-4, DOI: 10.1109/NCVPRIPG.2013.6776195.

[16]S. Deivalakshmi, K. Chaitanya and P. Palanisamy, ”Detection of table structure and content extraction from scanned documents,” 2014 International Conference on Communication and Signal Processing, 2014, pp. 270-274, DOI: 10.1109/ICCSP.2014.6949843.

[17]Alaei, Alireza, Umapada Pal, P. Nagabhushan. ”A new scheme for unconstrained handwritten text-line segmentation.” Pattern Recognition 44, no. 4 (2011): 917-928.

[18]A. Sheshkus, A. Ingacheva, V. Arlazarov, and D. Nikolaev, ”HoughNet: Neural Network Architecture for Vanishing Points Detection,” 2019 International Conference on Document Analysis and Recognition (IC- DAR), 2019, pp. 844-849, DOI: 10.1109/ICDAR.2019.00140.

[19]H. All e´ritie`re, F. Cloppet, C. Kurtz, J. Ogier, and N. Vincent, ”A Document Straight Line Based Segmentation for Complex Layout Extraction,” 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), 2017, pp. 1126-1131, DOI: 10.1109/IC- DAR.2017.186.

[20]A. Yuan, G. Bai, P. Yang, Y. Guo, and X. Zhao, ”Handwritten English Word Recognition Based on Convolutional Neural Networks,” 2012 International Conference on Frontiers in Handwriting Recognition, 2012, pp. 207-212, DOI: 10.1109/ICFHR.2012.210.

[21] Kasturi, Rangachar, Lawrence O’gorman, and Venu Govindaraju. ”Document image analysis: A primer.” Sadhana 27, no. 1 (2002): 3-22.

[22] Gupta, Maya R., Nathaniel P. Jacobson, and Eric K. Garcia. ”OCR binarization and image pre-processing for searching historical documents.” Pattern Recognition 40, no. 2 (2007): 389-397.

[23]Vijayashree, C. S., & Vasudev, T. AModel TO CONVERT WAVE–FORM-TEXT TO LINEAR-FORM-TEXT FOR BETTER READABILITY BY OCRS.

[24]Sharda, V., & Kishan, A. C. (2009). Skew Detection and Correction in Scanned Document Images (Doctoral dissertation).

[25]Abuhaiba, I. S. (2003). Skew correction of textural documents. Journal of King Saud University-Computer and Information Sciences, 15, 73-93.

[26]Cao, Y., Wang, S., & Li, H. (2003). Skew detection and correction in document images based on straight-line fitting. Pattern Recognition Letters, 24(12), 1871-1879.

[27]Singh, C., Bhatia, N., & Kaur, A. (2008). Hough transform-based fast skew detection and accurate skew correction methods. Pattern Recognition, 41(12), 3528-3546.

[28]Kumar, D., & Singh, D. (2012). The modified approach of hough transforms for skew detection and correction in documented images. International Journal of Research in Computer Science, 2(3), 37.

[29]Dong, J. X., Dominique, P., Krzyyzak, A., & Suen, C. Y. (2005, August). Cursive word skew/slant corrections based on Radon transform. In Eighth International Conference on Document Analysis and Recognition (ICDAR'05) (pp. 478-483). IEEE.

[30]Raducanu, B., Boiangiu, C. A., Olteanu, A., Ștefãnescu, A., Pop, F., & Bucur, I. (2011, May). Skew Detection Using the Radon Transform. In Proceedings CSCS-18, The 18th International Conference on Control Systems and Computer Science (pp. 653-657).

[31]Rosner, D., Boiangiu, C. A., Zaharescu, M., & Bucur, I. (2014, May). Image skew detection: A comprehensive study. In Proceedings of IWoCPS-3, The Third International Workshop On Cyber-Physical Systems, Bucharest, Romania.

[32]Shah, L., Patel, R., Patel, S., & Maniar, J. (2014). Skew detection and correction for Gujarati printed and handwritten characters using linear regression. International Journal, 4(1).

[33]Del Ninno, E., Nicchiotti, G., & Ottaviani, E. (1997, September). A general and flexible deskewing method based on generalized projection. In International Conference on Image Analysis and Processing (pp. 632-638). Springer, Berlin, Heidelberg.

[34]Bafjaish, S. S., Azmi, M. S., Al-Mhiqani, M. N., Radzid, A. R., & Mahdin, H. (2018). Skew detection and correction of Mushaf Al-Quran script using hough transform. International Journal of Advanced Computer Science and Applications, 9(8), 402-409.

[35]Ali, A. A. A., & Suresha, M. (2017). A novel approach to correction of skew at document level using an Arabic script. Int J Comput Sci Inf Technol, 8(5), 569-573.

[36]Boudraa, O., Hidouci, W. K., & Michelucci, D. (2020). Using skeleton and Hough transform variant to correct skew in historical documents. Mathematics and computers in simulation, 167, 389-403.

[37]Subrahmanyam, M. S. L. B., Kumar, V. V., & Reddy, B. E. (2018). A new algorithm for skew detection of Telugu language document based on principle-axis farthest pairs quadrilateral (PFPQ). International Journal of Image, Graphics and Signal Processing, 10(3), 47.

[38]Bezmaternykh, P. V., & Nikolaev, D. P. (2020, January). A document skew detection method using fast Hough transform. In Twelfth International Conference on Machine Vision (ICMV 2019) (Vol. 11433, p. 114330J). International Society for Optics and Photonics.