SN Computer Science (2023) 4:253
https://doi.org/10.1007/s42979-023-01693-5

ORIGINAL RESEARCH

Computer Vision Based Automatic Margin Computation Model for Digital Document Images

Abhijit Guha[1,2] · **Debabrata Samanta**[1,3] · **Sandeep Singh Sengar**[4]

Received: 11 October 2022 / Accepted: 10 January 2023 / Published online: 7 March 2023

Abstract

Margin, in typography, is described as the space between the text content and the document edges and is often essential information for the consumer of the document, digital or physical. In the present age of digital disruption, it is customary to store and retrieve documents digitally and retrieve information automatically from the documents when necessary. Margin is one such non-textual information that becomes important for some business processes, and the demand for computing margins algorithmically mounts to facilitate RPA. We propose a computer vision-based text localization model, utilizing classical DIP techniques such as smoothing, thresholding, and morphological transformation to programmatically compute the top, left, right, and bottom margins within a digital document image. The proposed model has been experimented with different noise filters and structural elements of various shapes and size to finalize the bilateral filter and lines and structural elements for the removal of noises most commonly occurring due to scans. The proposed model is targeted towards text document images and not the natural scene images. Hence, the existing benchmark models developed for text localization in natural scene images have not performed with the expected accuracy. The model is validated with 485 document images of a real-time business process of a reputed TI company. The results show that 91.34% of the document images have conferred more than 90% IoU value which is well beyond the accuracy range determined by the company for that specific process.

Keywords Text localization · Margin detection · Digital image processing · Computer vision

This article is part of the topical collection "Cyber Security and Privacy in Communication Networks" guest edited by Rajiv Misra, R K Shyamsunder, Alexiei Dingli, Natalie Denk, Omer Rana, Alexander Pfeiffer, Ashok Patel and Nishtha Kesswani.

Abhijit Guha, Debabrata Samanta and Sandeep Singh Sengar have contributed equally to this work.

✉ Sandeep Singh Sengar
SSSengar@cardiffmet.ac.uk

Abhijit Guha
abhijit.guha@res.christuniversity.in

Debabrata Samanta
debabrata.samanta369@gmail.com

[1] Department of Data Science, CHRIST (Deemed to be) University, Bengaluru, Karnataka, India

[2] R &D Scientist, First American India Private Limited, Bengaluru, Karnataka, India

[3] Department of Computational Information Technology, Rochester Institute of Technology , Germia Campus, 10000 Prishtina, Kosovo

[4] Department of Computer Science, Cardiff Metropolitan University, Cardiff CF5 2YB, UK

Introduction

Historically, margins have been used as a method to layout text within a document. The use of it dates back to ancient Egypt when people used papyrus scrolls to organize writings, and margins were the visual mark indicating the end of one line and the beginning of another. Since the invention of the codex, the need for margin to distinguish text blocks were no more relevant. However, instead of becoming antiquated, it took on a new role. The margin provided a visual aesthetic to the text and allowed the reader to put down the notes and commentary within the blank space. It is ubiquitous in the twenty-first century to organize the texts into digital form, and the margin remains and is utilized for placing signatures, stamping, notes, etc. In some business processes, the margin has become indispensable for the above reasons. One such real-time obligation for margin is observed in the TRS, which is the primary motivation of the present research (see Fig. 2).

Recordation of legal instruments in a county recording office that is a public registry is an act of constructive public

notice of ownership to the subsequent purchaser, creditor, or mortgagers. The recordation itself does not determine the title but provides a framework for the legal system to do so during any future litigation. Various TI companies provide the TRS to their customers, facilitating the recording. The state statute establishes the rules of recording. The rules differ from county to county. The documents to be recorded must comply with local and state requirements. Some of the standard regulations most of the counties in the USA follow are as follows;

- The documents must be notarized.
- There must be county-specified margins on the top, left, right, and bottom of every document's first and last page.
- If the specified margin is absent, a cover sheet must be attached for a recording stamp.
- Original signatures must be present on all instruments.

Before the recording, the service providers verify the rules manually. Human intervention makes the validation time-consuming and error-prone, which in turn increases the cost of recording. An intelligent system to automate the verification is the need of the hour for the TRS providers.

Numerous studies related to text localization and recognition have been conducted in the recent past to read text from various scenes and videos for content-based analysis. It is an essential prerequisite for OCR of a digital text document. Finding out the text from the image document is the first step for any OCR product to extract the characters from the text. Applications to assist visually impaired individuals with the surrounding entities to give them a certain autonomy have been around for some time. Reading license plates and automatic navigation are some of the other applications of text localization. To our knowledge, no previous study to date examined the computation of the margin within a digital image document to automate the recording process. A whole range of different approaches is administered to detect text areas within a document or a natural scene in the past decade.

In the present study, we have adopted a classical design to traditional image processing techniques for detecting the text segments within the document image after correcting the skewness (See Fig. 1). DIP operations such as noise removal, image binarization followed by morphological transformation are carried out sequentially. Additionally, to eliminate the vertical or horizontal prominent noise on the edges of the documents commonly noticed in scanned document images, vertical and horizontal structure detector kernels are interjected. After detecting all the text areas within the document, the bounding regions are merged into one frame to accommodate all the text within the minimum rectangular bounding box. The predicted text region is then compared with the GT manually drawn area. Finally, IoU is calculated

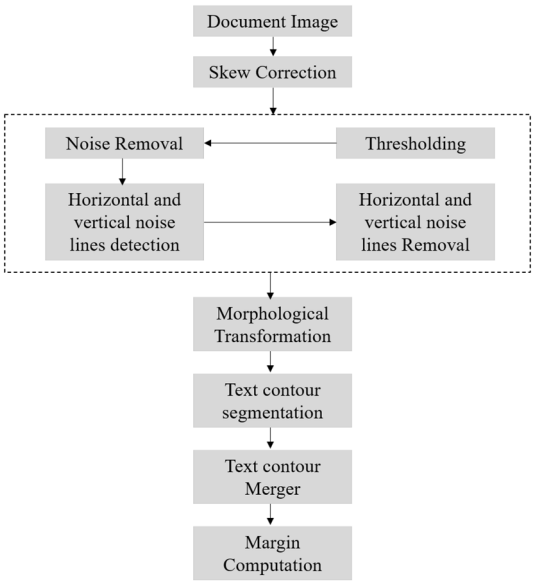


Fig. 1 AMCM process flow

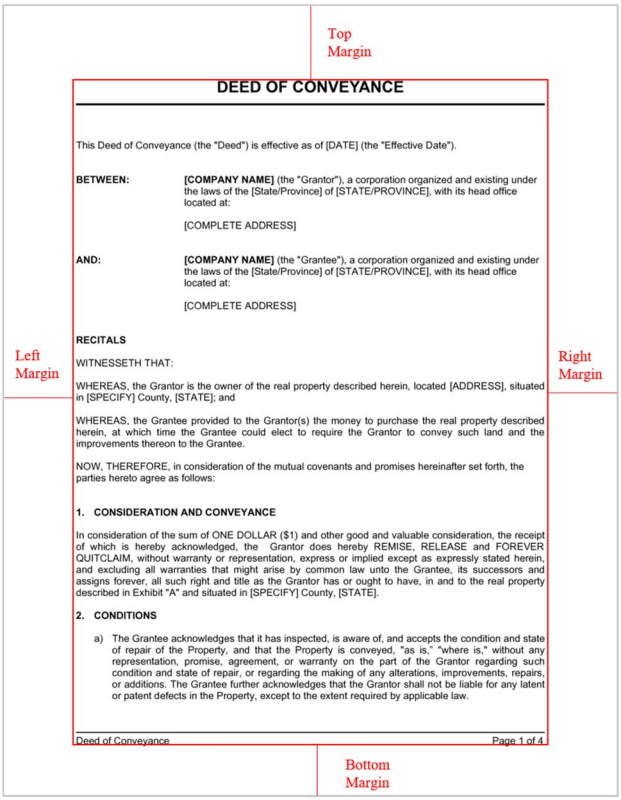


Fig. 2 An example document image with a bounding box enclosing the text region; Top, left, bottom and right margins are labelled

from both the bounding boxes to gauge the accuracy of the algorithm. The margin is calculated by subtracting the computed rectangle from the original image. IoU for 485

document images are computed using the model, and more than 91.34% of the computed IoU are found to be more than 90% accurate. The margins are initially calculated in pixels and converted to inches for consumption by the client applications. The objective is to devise a model that automatically computes the top, left, right, and bottom margins of a given page of a document image by locating the text area (foreground) within the image background without human intervention. Based on the automatically computed margin, the consumer application can verify the county requirement and take subsequent workflow or process flow decisions.

Motivation

- TRS providers spend a significant amount of time and effort checking the prerequisites of the recording rules. This human cognition-dependent step increases the cost of the service. In the dawn of the technological revolution, there is a dire need for research to seek the possibility of real-time, intelligent automation in this regard.
- Several sophisticated machine learning-based techniques are explored in the past to solve the text localization problem for various real-time applications. Still, little emphasis has been paid to the classical DIP and computer vision techniques which can effectively be utilized in scenarios like this.
- Besides the fact that the chance of error and risk and processing time increases due to human validation of such recording prerequisite examinations, it is a classic example of misuse of the potential of human cognition. The tasks are monotonous and repetitive to be performed by a dedicated expert system in place of human associates.

Contribution

- The proposed model is validated in a real-time environment with actual production digital documents, and the presented results are for consumption for business decisions.
- Although the research has been conducted with digital documents from TRS, the model can be utilized for margin computation for any digital documents irrespective of any domain.
- The model produced as high as 90% IoU score for more than 91% documents including samples with significant shot noise and edge noises occurring due to poor scanning.

Text localization is a prerequisite for the information extraction and solution approaches have been categorized as Region based, Texture based and Hybrid [50]. A host of artificial neural network and machine learning based architectures have been proposed that have proved the efficacy over the time. ICDAR has been the standard benchmark data set for all the past experiments. However, such sophisticated methods have proved to be less effective for a much simpler task of text localization for the scanned digital image documents for margin detection. The machine learning based methods have turned out to be time consuming and costly. Our proposed method depends on simpler image processing methods without the need of training hazards and address all sorts of digital scanned documents with variety of noise.

Paper Organization

The remainder of the paper is divided into five sections. A thorough literature survey of different techniques of text localization and its application is presented in "Background and Related Work". The dataset and the description of the methods used are briefed in "Materials and Methods", followed by the experiment steps and the results obtained in "Experiment". Finally, in "Results and Discussion" and "Conclusion with Future Scope", the insights obtained from the result are discussed, followed by the conclusive remarks and future scope.

Background and Related Work

'Text' being one of the most expressive mediums of communication, can be primarily embedded in three different digital sources; scanned documents, random images, and videos [38, 43, 47]. Increased attention to detect and recognize text from the sources described above is seen in recent times. Text localization is a sub-problem of detection which focuses on locating the region where the text data is present within a given image rather than recognizing its semantics [23–27].

The literature in this regard is divided based on the source where the text is to be localized. The advance in computer vision and pattern recognition also has made it more feasible to address different challenges faced during text detection and localization [23, 28]. Further categorization of the literature can be perceived based on the technology, such as deep learning-based, statistical and CV-based.

Text Localization in Document Image

Khan et al., in their study proposed a hybrid technique for localizing text elements from both document and scene images. They applied Morphological operations to segregate the foreground objects that is text from the backgrounds.

They adopted MLP approach to classify the text regions and non text regions using statistical features. The proposed method achieved 86.38% accuracy for text region isolation [29].

Nikitin et al. proposed a two step architecture to detect the word level bounding box followed by text region identification using classic computer vision techniques. Their proposed model outperformed the other state-of-the-art techniques of text localization for document images [30].

Nagaoka et al. used R-CNN object detection method for text localization. They introduced an additional layer called merge layer that enables multiple region of interest more effectively than the standard R-CNN network [31]. It is a primary prerequisite to detect and segment text regions within a document image for transcription. The task is even more difficult when the text is handwritten. Carbonell et al. in their study [44] proposed the technique of full page transcription after text area segmentation.

Neumann et al. proposed an end-to-end unconstrained text localization and recognition method for text region detection. The method is a deviation from the common assumption of region-based approaches of connected component analysis [49].

Materials and Methods

The proposed model is designed to compute the margin by locating the text area within the document image. Once the text areas are segmented, the margin is calculated by subtracting the text area from the remaining document. Although various state-of-the-art text localization techniques are adopted in the recent past, we have chosen multiple classical digital image processing techniques for text localization. Having dealt with digital text document images, the complexity of text localization from the natural scene needed no special attention during the experiments.

Table 1 Structure of the ground truth bounding region data collected trough VIA

	x	y	h	w
File1.png	94	100	981	641
File2.png	87	105	700	356
File3.png	58	134	789	567
File4.png	45	78	678	345
.	.	.	.	.

Data Set

Thirteen different types of digitized, real-time, recordable, multi-page, legal instruments that a reputed TI company uses for the recording are collected and used as samples for the experiment. The document types and the distribution of the total number of pages considered are as shown in Fig. 3.

Every page of the document is regarded as an independent recordable instrument to improve the study's rigor. However, the real-time recordation considers the first page of every document as a recordable instrument, and the recording office provides stamps only on the first page. Four thousand five hundred sixty-seven individual document pages are annotated with the margins manually. The system predicted margins for every document page are programmatically compared with the manually annotated margins.

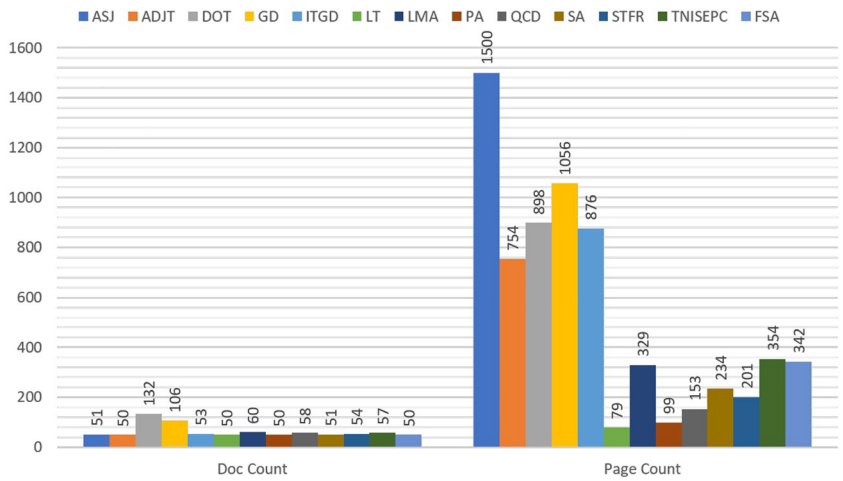


Fig. 3 Test sample count and page count across thirteen different title insurance document types that are instruments for recording

Annotation

The model identified bounding box enclosing the text area within the image document is compared with the manually drawn bounding box. It is not expected to get the exact convergence of the both but convergence beyond a certain predetermined threshold can be considered for the evaluation of the model. Every single page of the documents were considered independent document image and the bounding box excluding the margin of the page was manually drawn using VIA [1, 2], an open source image annotation tool developed by VGG [3]. The annotated bounding box data for every image is saved in the below tabular format (see Table 1).

As described in Fig. 4, (x,y) denotes the top left pixel coordinate of the drawn rectangle and h and w denotes the height and width of the rectangle enclosing the text respectively. H and W are the original height and width of the image.

The human annotated data is compared with the model annotated data and IoU is calculated from the coordinates of both GT and the predicted coordinates. The model is built upon the below algorithm and the steps are described in the following sections.

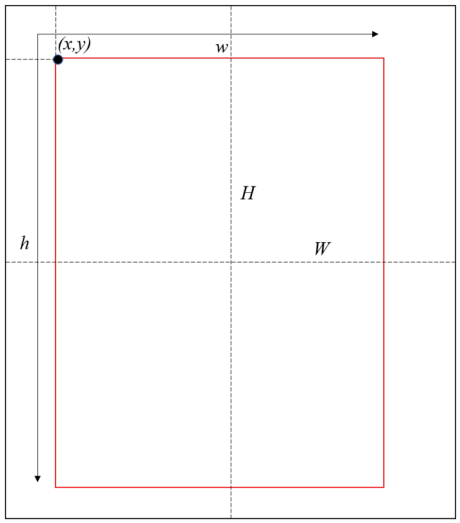


Fig. 4 Image coordinate system and the margin calculation within that reference

Skew Correction

Digital documents are often scanned from the physical copies, and many a time, the orientation of the text within a document is not appropriately aligned. It is an essential preprocessing step to deskew the text within the image to identify the margin space within a digital textual document image (as shown in Fig. 5).

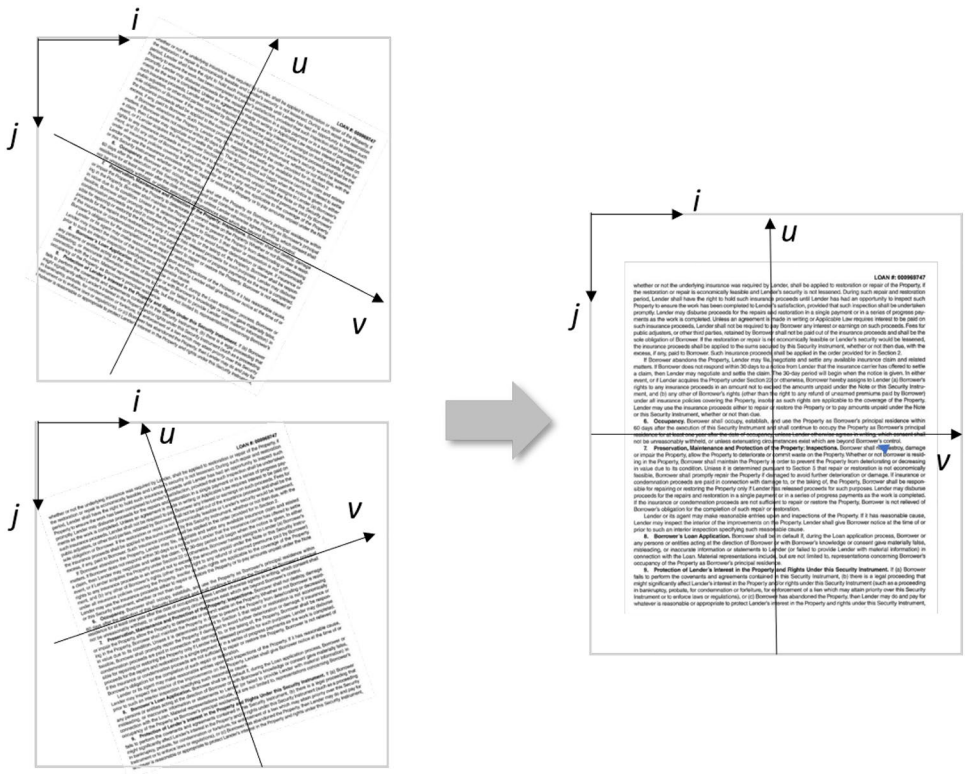


Fig. 5 Skew correction of a rotated document image

There are predominantly four techniques that are used for shew correction of text within a document image; Projection Profile, Hough Transform, Nearest Neighbour Clustering and Fourier Transform are explored and experimented in a number of different variations [4].

Gray-Scale Conversion

Digital document images, like other digital images, are usually represented in three-color channels, namely Red, Green, and Blue, also known as RGB color space [6]. However, colors seldom have much significance in the applications of digital image processing or computer vision algorithms. Gray-scale or monochrome representations are frequently used successfully by different descriptors that reduce the complexity and computational effort. Applications concerned with text data within digital images have little importance with the color space as the single-channel retains all the necessary information. In the present study, the interest is to localize the text regions within the document, which can be accomplished through a gray-scale image. In this pre-processing step, an RGB to the gray-scale conversion function κ is applied to a color image in $\mathbb{R}^{n\times m\times 3}$ space to convert it to a $\mathbb{R}^{n\times m}$ representation [5] where the pixel values are within 0 representing the strongest intensity and 255, the weakest intensity of the contrast (Fig. 6). Subsequent processing steps are performed on the gray-scale version of the document image.

The three most commonly used color space conversion methods are based on Lightness, Average and Luminosity. Lightness based conversion averages the most prominent and the least prominent colors to represent the gray-scale pixel value.

$$\kappa_{Lightness} = \frac{max(R,G,B) + min(R,G,B)}{2} \tag{1}$$

The average method simply averages the RGB pixels of the colored image.

$$\kappa_{Average} = \frac{R+G+B}{3} \tag{2}$$

The luminosity method is a little more sophisticated weighted average method that accounts for the human perception of the color intensity.

$$\kappa_{Luminosity} = w_1 \cdot R + w_2 \cdot G + w_3 \cdot B, \tag{3}$$

where $w_1 = 0.21$, $w_2 = 0.72$ and $w_3 = 0.07$. As human eye is most sensitive towards the green color, the highest weight is assigned to the green channel. Luminosity based gray-scale conversion has been used for converting the digital documents to gray-scale, monochrome images in the present study.

Denoising

Noise is an unavoidable occurrence during the capture or transmission phase of a digital image that degrades the image quality and poses hindrances during image processing. Digital text images are no exception to this side-effect [7, 8]. Digital text documents are often printed and scanned multiple times before storing them in the digital archive. A common occurrence of noise is the dust particles present in the scanner or the printer screen, causing the edge or marginal noise within a digital text document (see Fig. 7). These noises make the text localization difficult within a digital document and frequently results in false detection of the noise area.

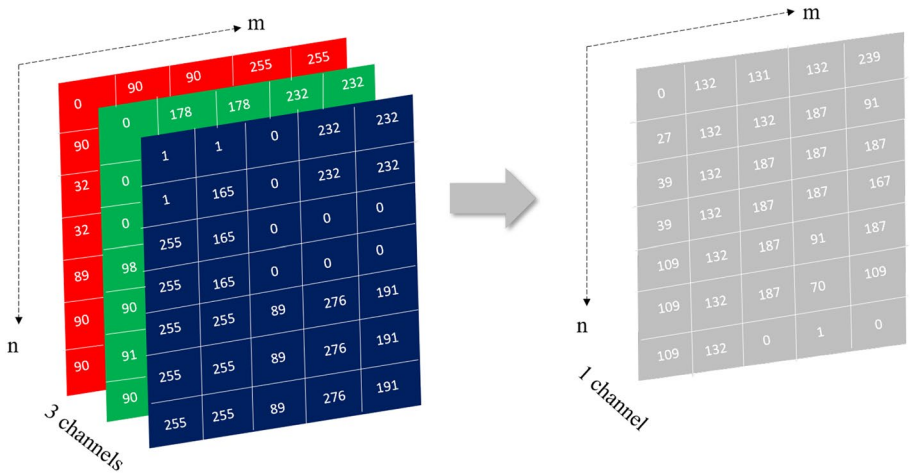


Fig. 6 RGB (multi chromatic channel) to gray scale (mono chromatic channel) conversion

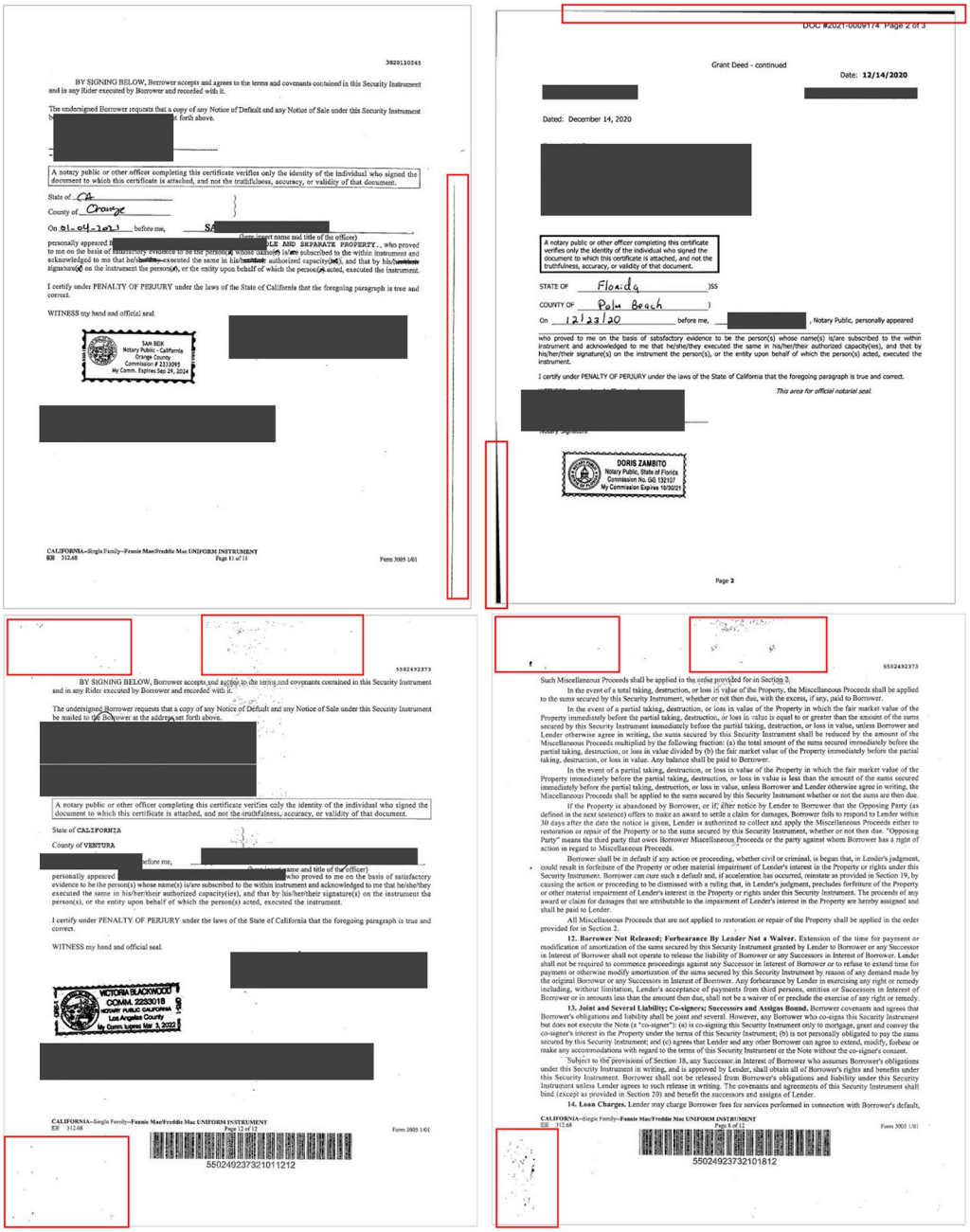


Fig. 7 Scanned image documents with occurrences of shot noise and edge noise

Below types of noise are predominantly observed in a digital image.

- Gaussian noise: Occurs during image acquisition mostly due to noise in the sensors.
- Salt and pepper noise: Brighter region of an image with dark pixels and the darker regions with bright pixels.
- Shot noise: Shot noises are dominant in the brighter regions of the image and often follow Poisson distribution.
- Quantization noise: Occurs due to sampling the image to discrete levels and follow uniform distribution.

Filtering techniques like mean, median, Gaussian, Bilateral, and Weiner filtering are standard and frequently used for

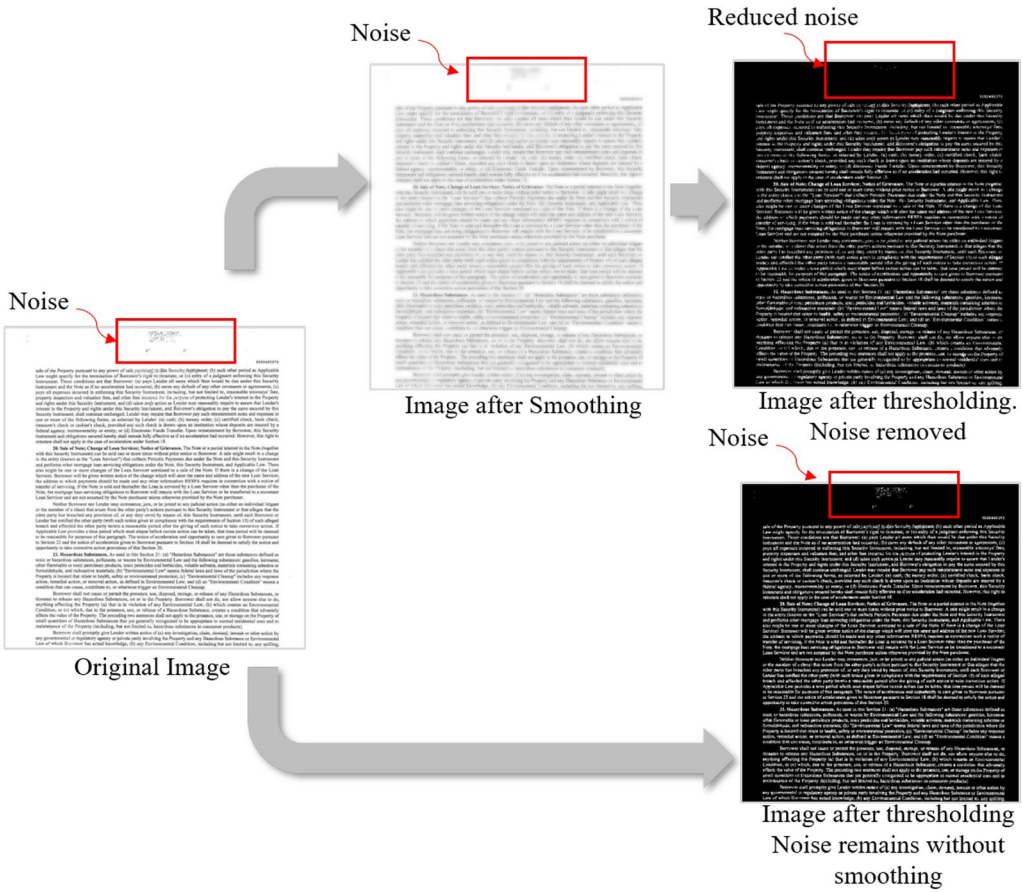


Fig. 8 Effect of smoothing using bilateral filter on scanned document image with shot noise. After binarization the noise is negligible for smooth image

noise removal and smoothing for subsequent processing. In the present study, bilateral filtering effectively removes the shot noise and edge noise occurring due to the scan of the document (as shown in Fig. 8).

Adaptive Thresholding

Document images concerning the recording process have primarily two segments within the image; the background, which is of lighter intensity, and the foreground text, which is of darker intensity [15, 16]. It is sufficient to segment the images into these two groups of pixels to segregate the text area's background and text area. As the text regions within the gray-scale image can have various intensity levels of gray, data-driven adaptive thresholding (Otsu's method) has been adopted for the binarization task [9, 12–14].

Otsu's binarization method identifies and returns a single intensity threshold for a given image to represent the image in two classes namely foreground (the pixels representing textual elements) and background (the pixels representing the empty section of the canvas). The optimal threshold is determined by minimizing within class variance and maximizing the between class variance through an exhaustive search algorithm.

$$\sigma_w^2(t) = w_0(t)\sigma_0^2(t) + w_1(t)\sigma_1^2(t) \tag{4}$$

w_0 and w_1 are t threshold separated probabilities and σ_0^2, σ_1^2 are variances of two classes respectively. The class probabilities are computed from the bins (L) of the histogram.

$$\begin{aligned} w_0(t) &= \sum_{i=0}^{t-1} pi \\ w_1(t) &= \sum_{i=t}^{L-1} pi \end{aligned} \tag{5}$$

Considering between class variance maximization is equivalent to minimizing within class variance,

$$\begin{aligned} \sigma_b^2(t) &= \sigma^2 - \sigma_w^2(t) \\ &= w_0(\mu_0 - \mu_T)^2 + w_1(\mu_1 - \mu_T)^2 \\ &= w_0(t)w_1(t)[\mu_0(t) - \mu_1(t)]^2 \end{aligned} \tag{6}$$

The class means are represented as

$$\mu_0(t) = \frac{\sum_{i=0}^{t-1} ip(i)}{w_0(t)}$$

$$\mu_1(t) \frac{\sum_{i=t}^{L-1} ip(i)}{w_1(t)} \tag{7}$$

$$\mu_T = \sum_{i=0}^{L-1} ip(i)$$

After the binarization, the gray-scale image is considered fit for a series of morphological transformation operations to remove other horizontal and vertical structured edge noise arising from the scanning (Fig. 6) and localize the foreground text features.

Morphological Transformation

MT is a series of non-linear mathematical operations on the morphology or shape of features in a digital image [19, 20]. These operations take the relative ordering of the neighboring pixel intensities into account and do not depend on the absolute pixel intensity; hence, the functions are best suited for a binary image. However, there are mathematical variations of MT available for gray-scale images. Although MT is usually considered to remove the imperfections occurring due to the binarization of an image, the present study utilizes the transformation not to improve the feature (foreground text) prominence but to approximate the foreground discovery.

A small image template called the SE or **Kernel** is convoluted over the original binary image to probe the presence or absence of certain shapes or structures within the image [17]. During the convolution of the kernel over the original image, the kernel is said to fit the image if each pixel with intensity 1 of the SE matches exactly with the pixel intensity of the original image and is said to hit if at least one pixel of the kernel set to 1 matches with that of the larger image.

Let I be the binary image in Euclidean space E and K is the kernel.

Erosion

The erosion of I by the kernel K produces an output image with 1's in the origin of the kernel at which K fits I [19]. It is denoted by

$$I \ominus K = \{z \in E | K_z \subseteq I\} \tag{8}$$

where K_z is the translation of K by vector z i.e. $K_z = \{b + z | b \in B\}, \forall_z \in E$.

Erosion reduces the boundary of regions of the white pixels or the foreground pixels (pixels representing the text in the digital document in the present case). The gaps within the regions holding the foreground pixels are enlarged [18, 21].

Dilation

The dilation of I by the kernel K produces an output image with 1's in the origin of the kernel at which K hits I. It is denoted by

$$I \oplus K = \bigcup_{b \in B} I_b \tag{9}$$

Dilation increases the boundaries of the regions by adding pixels to the foreground [19]. It improves or enhances the features.

Opening

Opening is a compound morphological operation represented by erosion of I by K, followed by dilation of the resultant image by K [19] denoted as,

$$I \circ K = (I \ominus K) \oplus K \tag{10}$$

This operation opens up the gap between two objects connected by a thin layer of pixels. The surviving pixels after the erosion are restored to the original size by dilation.

Closing

Closing is the reverse compound operation of opening which applies erosion on the resultant image obtained by the dilation operation of I by K [19], represented as

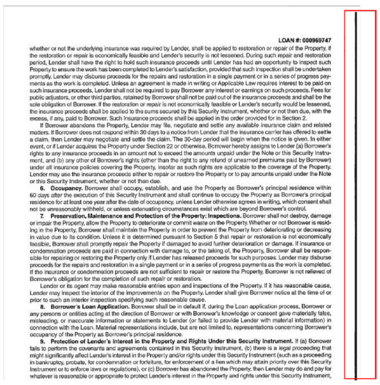
$$I \bullet K = (I \oplus K) \ominus K \tag{11}$$

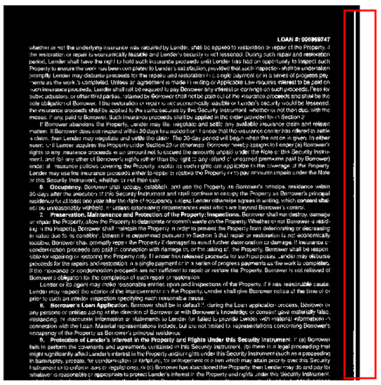
It helps closing small gaps or holes within a region of binary image [22].

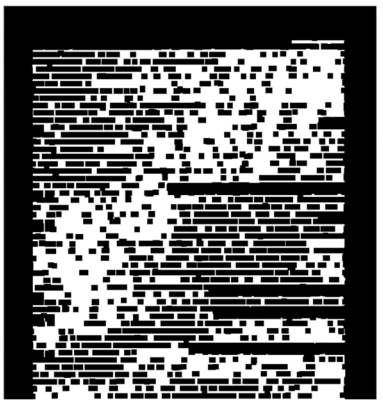
Primary objective of the applying MT in the experiment is to locate the text region within a document image with high accuracy so that the non text regions that are present in the foreground pixels due to noise can be excluded as much as possible. After applying the combination of the aforesaid MT, we have been able to detect and remove most of the noises with precision as shown in the below Fig. 9.

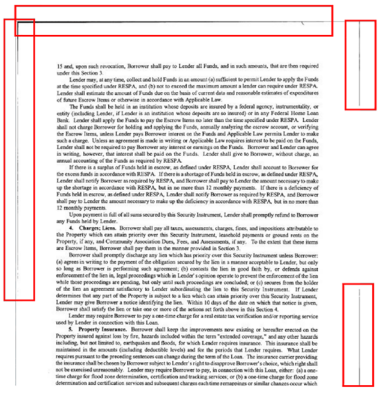
Experiment

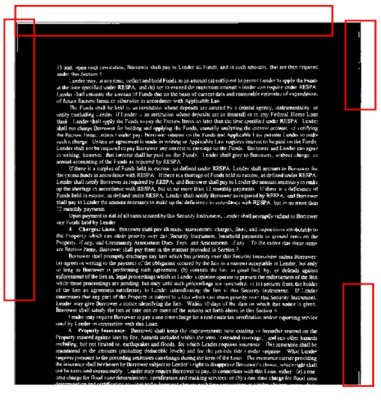
AMCM is validated using 485 various document image pages collected from a real-time TRS provided by a reputed TI company [46, 48]. Every page of the document is considered to be an independent image candidate for margin computation. Every image is manually annotated using VIA annotator. A single rectangle is drawn to

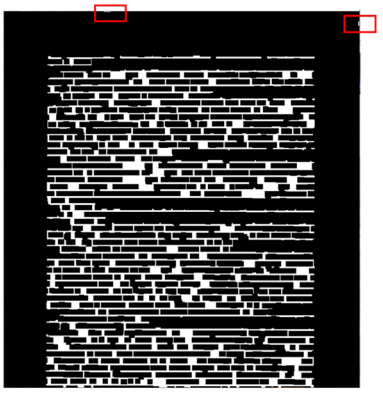


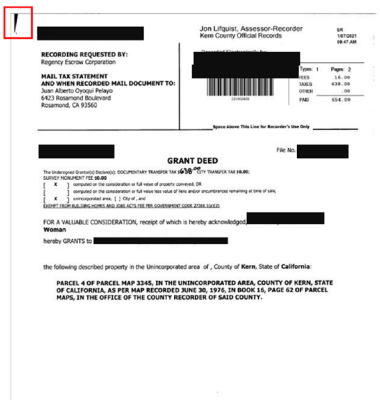


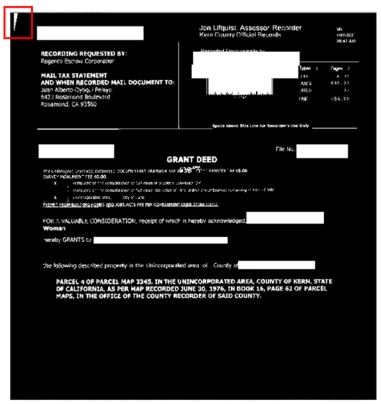


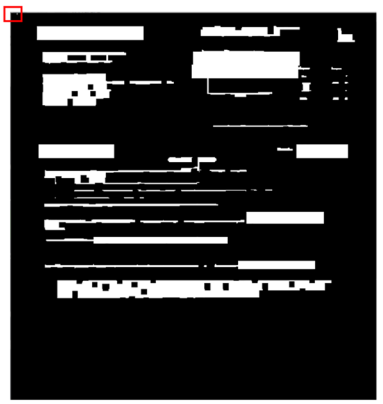


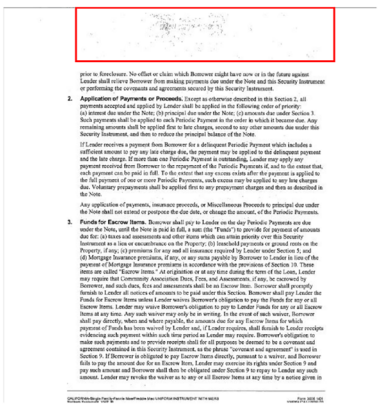


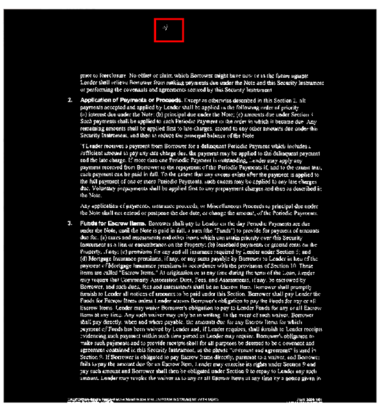












◂**Fig. 9** Scanned images with significant shot and edge noise in the first column. The second column shows the noise status after smoothing and binarization and the third column shows the impact of morphological operations. The noise become insignificant and the text area becomes prominent for localization

localize all the text regions within the document image. Finally the document images are passed through AMCM and the predicted rectangle is finally compared with the GT using the Intersection over Union (IoU) method.

Four different de-noising filters namely Gaussian filter, Mean filter, Median filter and Bilateral filter are evaluated by comparing the average IoU obtained (Table 2 below). Bilateral filter has obtained significantly higher IoU over the other filters.

Horizontal and vertical line kernels of different sizes with respect to the height (h) and width (w) of the image are considered and obtained IoU's are compared. Below table shows that the IoU obtained for the line kernel of size 0.05 times the h and w received the best IoU (Table 3).

The proposed classical DIP approach for margin detection is compared with the state-of-the-art object detection models YOLOv4 (You Only Look Once version 4) and Mask R-CNN (Regions with mask convolutional neural network). The same data set used in the experiment mentioned above is used. 485 document images were split into training and testing sets with 80 and 20 percent, respectively. Additionally, we have also utilized state-of-the-art OCR technologies like Google Vision, Azure Cognitive OCR, and Tesseract for text localization and calculated the margins from the word boundaries retrieved from the OCR output. For all the methods, we have calculated the average IoU and compared it with the average IoU of the proposed model. The comparison is shown in Table 4.

Evaluation Metric

IoU is a popular evaluation metric used for an OD to measure the accuracy of the localization of the detected object. As long as, there is a GT bounding box (drawn by a human) available to be compared with a machine predicted bounding box, the IoU can be utilized to measure the accuracy of the machine prediction (Fig. 10 depicts the IoU calculation). It is extremely unlikely that the predicted bounding box and the GT bounding box will exactly overlap with each other, pixel by pixel. However, the higher the overlap better the prediction.

Table 2 IoU comparison with different noise filters

Filter	Average IoU
Gaussian	0.9346
Mean	0.9583
Median	0.8209
Bilateral	0.9732

This is also known as Jaccard similarity coefficient, a statistic that measures the similarity between two finite sample sets S_1 and S_2 using the below set operation.

$$J(S_1, S_2) = IoU = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{12}$$

$$J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1| + |S_2| - |S_1 \cap S_2|} \tag{13}$$

By design, $0 \leq J(S_1, S_2) \leq 1$. The GT bounding box is represented by S_1 and predicted bounding box by S_2 (as depicted in Fig. 11). There is no absolute value to determine the accuracy as good or bad. It depends on the specific OD problem. In the present study, our aim is to detect the minimum area covering the foreground text region and the threshold of 0.9 that corresponds to 90% overlapping region is considered to be a very close or acceptable accurate prediction by the algorithm.

The most common metrics used in object detection are AP and mAP. These metrics are calculated combining the metrics for object classification as well as object localization. In the present proposed study, the interest is only to localize the object (the text region) and not the classification.

Table 3 IoU comparison with different noise filters

Kernel size (times h and w)	Average IoU
0.2	0.9346
0.171	0.9383
0.142	0.9309
0.125	0.9454
0.111	0.9309
0.1	0.9429
0.091	0.9538
0.083	0.9232
0.077	0.9013
0.071	0.9567
0.062	0.9512
0.063	0.9398
0.058	0.9622
0.056	0.9732
0.052	0.9701
0.05	0.9706
0.048	0.9632
0.045	0.9672
0.043	0.9332
0.042	0.9522
0.04	0.9519

Algorithm 1 Margin Computation Algorithm

Inputs

- Image I; $Height = h$;$Width = w$
- GT bounding box coordinates; $x_{GT}, y_{GT}, h_{GT}, w_{GT}$;

Steps

1. $I_{grayscale} = ToGrayscale(I)$
2. $I_{filtered} = BilateralFilter(I_{grayscale})$
3. $I_{binary} = Binarization(I_{filtered})$
   $I_{inverse} = I_{binary}^{-1}$.
4. Initialize Horizontal Kernel HK;
   $I_{transformed} = Erode(I_{inverse}, HK)$;
5. Initialize Vertical Kernel VK;
   $I_{transformed} = Dilate(I_{transformed}, VK)$;
6. Initialize Rectangular Kernel $RK_{(7X7)}$;
   $I_{transformed} = Closing(I_{transformed}, RK_{(7X7)})$
7. Initialize Rectangular Kernel $RK_{(3X3)}$;
   $I_{transformed} = Erode(I_{transformed}, RK_{(3X3)})$
8. $Contours = FindContours(I_{transformed})$;
   $c =$ count of bounding boxes;
   for **For** $i = 0$ to c
   (a) $X = x[i]$
   (b) $Y = y[i]$
   (c) $X_1 = x[i] + w$
   (d) $Y_1 = y[i] + h$
9. Merge the contours.
   $LM = min(X)$
   $RM = min(Y)$
   $TM = max(X_1)$
   $BM = max(Y_1)$
10. Convert pixels to inches.
    $RM_{inches} = (RM * 0.0104166667)$
    $LM_{inches} = (LM * 0.0104166667)$
    $TM_{inches} = (TM * 0.0104166667)$
    $BM_{inches} = (BM * 0.0104166667)$
11. Calculate IoU

IoU for every image is calculated for measuring the overall performance of the model. Figure 10 below depicts the IoU calculation. As it is quite unlikely to get the IoU value 1, an IoU threshold of 0.9 is considered as a threshold for a near-perfect prediction, and 10% error window is acceptable in the real-time scenario.

Results and Discussion

The IoU distribution is captured for 485 test image documents as shown in Fig. 12. The maximum and minimum IoU score attained by the algorithm is 0.99700 and 0.04384 respectively. The median IoU score is 0.96842 (see Table 5).

IoU values are rounded off to two decimal places to get a cumulative distribution pattern (Fig. 13). Out of 485 document images, we obtained IoU as 0.99 for 56, 0.98 for 110 and 0.97 for 100 documents which means that the system obtained an $IoU \geq 0.97$ for 54.84% of the observations. Empirical observation shows that the $IoU \geq 0.9$ are extremely accurate. Considering 0.9 as the threshold, we see from Table 6 is that 91.34% of the documents obtained an IoU beyond the threshold.

If the threshold is relaxed to 0.8 which as per IoU definition is fairly high accuracy of prediction, 97.93% of the documents fall within the positive prediction class. Based on the business need, the threshold can be moved up and down to decide the proportion of document to be passed through a human eyeballing.

We see from the result in Table 4 that the classical DIP technique outperforms all the state-of-the-art techniques in terms of detecting the text boundary and calculating the margin. However, accuracy is not the only advantage that is of prime concern in the study. YOLOv4 and Mask R-CNN needs heavy training and tagging cost in order to achieve respectable accuracy. The variation of the image documents is large and it is difficult to generalize a margin detection solution with a small set of sample training. On the other

Table 4 IoU comparison state-of-the-art text localization techniques

Method	Average IoU
YOLOv4	0.5739
Mask R-CNN	0.4908
Google OCR	0.8012
Azure Cognitive OCR	0.8112
Tesseract	0.7322
Proposed Method	**0.9732**

hand, the OCR techniques need two-step operations. First, get the OCR output, and second, calculate the margin from the token localized coordinates. This makes the solution heavily dependent on another solution and the cost factor is also included. The only open-source OCR solution that has been validated here has produced an IoU of 0.7322 which is considerably lower than that of the proposed method.

Conclusion with Future Scope

Despite the fact that advanced state-of-the-art AI and ML technologies are widely utilised in CV applications, the suggested model using fundamental computer vision techniques has empirically demonstrated to achieve a highly accurate detection of margins inside a digital image. More than 90% of 485 randomly selected digital documents from a real-time process of TRS of a reputed TI, where there were significant variations of noise present, our model was able to achieve 91.34% IoU. The result is encouraging for the adoption of the model in any business process with the requirement of margin detection across domains. The model uses classical approach of image binarization, color space conversion, Image inversion, Noise reduction and Morphological operations with three different kernels to accurately remove non text foreground pixels and detect and localize the textual foreground pixel within a single bounding box. Finally, the margin computation is a simple subtraction of the predicted bounding box containing the text region from the image height and width.

Future research could benefit from looking at digital papers from other industries, such as healthcare, tourism, and other BFSI. Although the model has been successfully tested on documents with prominent edge noises with significant vertical and horizontal structures, as well as shot noises caused by dirt in the scanner, documents with other types of noises, particularly those overlapping the document's margin, must be validated for accuracy. Furthermore, the study was conducted with documents that contained English language content, and the model must be tested with documents in other languages.

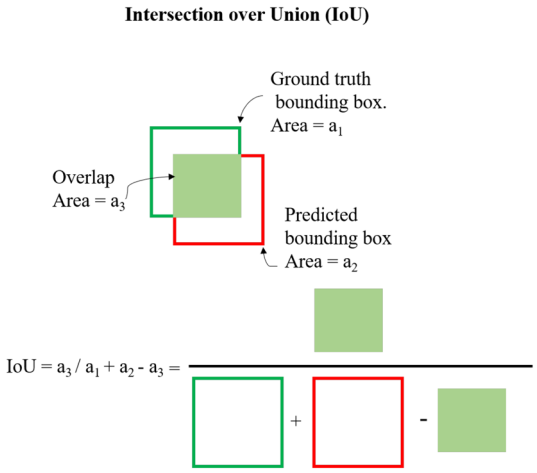


Fig. 10 Geometrical explanation of Intersection over Union

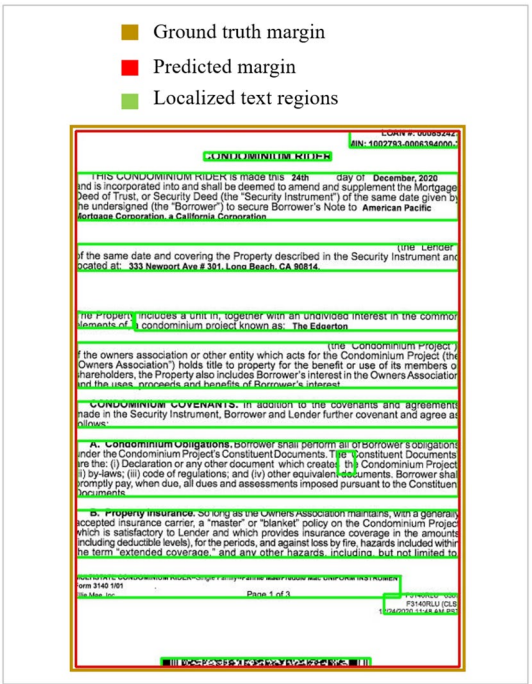


Fig. 11 Document Image with the predicted bounding box after merging the detected independent text blocks and the ground truth bounding box

Table 5 Summary of IoU distribution

Min	*Q*.1	*Median*	*Mean*	*Q*.3	*Max*
0.04384	0.93726	0.96842	0.94668	0.97894	0.99700

Fast R-CNN, Faster R-CNN, HOG, R-CNN, R-FCN, SSD, SPP-net, YOLO, etc. are some state-of-the-art supervised object detection algorithms based on deep neural

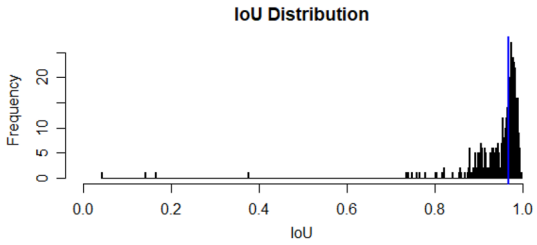


Fig. 12 IoU distribution with median indicated by the blue line

Table 6 Cumulative percentage of IoU score

IoU	Frequency	CF	Percentage	CP
0.98	110	110	22.6804	22.6804
0.97	100	210	20.6185	43.2989
0.99	56	266	11.5463	54.8453
0.96	54	320	11.1340	65.9793
0.93	26	346	5.3608	71.3402
0.94	23	369	4.7422	76.0824
0.95	22	391	4.5360	80.6185
0.91	20	411	4.1237	84.7422
0.9	19	430	3.9175	88.6597
0.92	13	443	2.6804	91.3402
0.88	10	453	2.0618	93.4020
0.89	10	463	2.0618	95.4639
0.82	3	466	0.6185	96.0824
0.86	3	469	0.6185	96.7010
0.74	2	471	0.4123	97.1134
0.76	2	473	0.4123	97.5257
0.8	2	475	0.4123	97.9381
0.04	1	476	0.2061	98.1443
0.14	1	477	0.2061	98.3505
0.16	1	478	0.2061	98.5567
0.37	1	479	0.2061	98.7628
0.75	1	480	0.2061	98.9690
0.78	1	481	0.2061	99.1752
0.84	1	482	0.2061	99.3814
0.85	1	483	0.2061	99.5876
0.87	1	484	0.2061	99.7938
1	1	485	0.2061	100.0000

network and computer vision, that can be further applied to the same data set and compared with the proposed accuracy.

Acknowledgements There is no funding available for this research work.

Declarations

Conflict of Interest The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

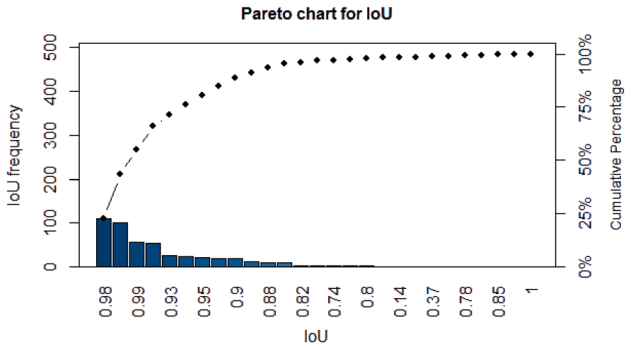


Fig. 13 Pareto chart for IoU distribution

Data availability The evaluation data that support the findings of this study are available on request from the corresponding author.

References

1. Dutta A, Gupta A, Zissermann A, VGG image annotator (VIA). http://www.robots.ox.ac.uk/vgg/software/via;2016.
2. Dutta A, Zisserman A. The VIA annotation soϕϕftware for images, audio and video. Proceedings of the 27th ACM International Conference on Multimedia. 2019;pp. 2276–2279.
3. Pizenberg M, Carlier A, Faure E, Charvillat V. Web-based configurable image annotations. Proceedings of the 26th ACM international conference on Multimedia. 2018;1368–1371.
4. Jundale TA, Hegadi RS. Skew detection and correction of Devanagari script using Hough transform. Proc Comput Sci. 2015;45:305–11.
5. Kanan C, Cottrell GW. Color-to-grayscale: does the method matter in image recognition? PLoS ONE. 2012;7(1): e29740.
6. Güneş A, Kalkan H, Durmuş E. Optimizing the color-to-grayscale conversion for image classification. SIViP. 2016;10(5):853–60.
7. Hambal AM, Pei Z, Ishabailu FL. Image noise reduction and filtering techniques. IJSR. 2017;6(3):2033–8.
8. Win N, Kyaw K, Win T, Aung P. Image noise reduction using linear and non-linear filtering technique. Int J Sci Res Publ. 2019;9(8):816–21.
9. Sengar SS, Mukhopadhyay S. 'Detection of moving objects based on enhancement of optical flow. Optik. 2017;145:130–41.
10. Gavaskar RG, Chaudhury KN. Fast adaptive bilateral filtering. IEEE Trans Image Process. 2018;28(2):779–90.
11. Sugimoto K, Kamata S-I. Compressive bilateral filtering. IEEE Trans Image Process. 2015;24(11):3357–69.
12. Goh TY, Basah SN, Yazid H, Safar MJA, Saad FSA. Performance analysis of image thresholding: Otsu technique. Measurement. 2018;114:298–307.

13. Nie F, Zhang P, Li J, Ding D. A novel generalized entropy and its application in image thresholding. Signal Process. 2017;134:23–34.
14. Sengar SS, Mukhopadhyay S. Moving object area detection using normalized self-adaptive optical flow. Optik. 2016;127(16):6258–67.
15. Ayala HVH, dos Santos FM, Mariani VC, dos Santos Coelho L. Image thresholding segmentation based on a novel beta differential evolution approach. Expert Syst Appl. 2015;42(4):2136–42.
16. Mlakar U, Potočnik B, Brest J. A hybrid differential evolution for optimal multilevel image thresholding. Expert Syst Appl. 2016;65:221–32.
17. Sreedhar K, Panlal B. Enhancement of images using morphological transformation. 2012; 2514 arXiv preprint arXiv:1203.2012
18. Brennan R. Quenching and morphological transformation in semi-analytic models and CANDELS. Mon Not R Astron Soc. 2015;451(3):2933–56.
19. Ashwitha K, Srikanth R. morphological background detection for enhancement of images. LAP LAMBERT Academic Publishing.2018
20. Jiménez-Sánchez AR. Morphological background detection and enhancement of images with poor lighting. IEEE Trans Image Process. 2009;18(3):613–23.
21. Bhatia G, Chahar V. An enhanced approach to improve the contrast of images having bad light by detecting and extracting their background. Int J Comput Sci Manag Stud. 2011;11(2):2231–5268.
22. Narasimhan K, Sudarshan CR, Raju N. A comparison of contrast enhancement techniques in poor illuminated gray level and color images. Int J Comput Appl. 2011;25(2):17–25.
23. Ye Q, Doermann D. Text detection and recognition in imagery: a survey. IEEE Trans Pattern Anal Mach Intell. 2014;37(7):1480–500.
24. Zhu Y, Yao C, Bai X. Scene text detection and recognition: recent advances and future trends. Front Comp Sci. 2016;10(1):19–36.
25. Yin X-C, Yin X, Huang K, Hao H-W. Robust text detection in natural scene images. IEEE Trans Pattern Anal Mach Intell. 2013;36(5):970–83.
26. Karatzas D. 'ICDAR 2013 robust reading competition', 2013 12th International Conference on Document Analysis and Recognition.2013; 1484–1493.
27. Karatzas D. 'ICDAR 2015 competition on robust reading', 2015 13th International Conference on Document Analysis and Recognition (ICDAR).2015;1156–1160.
28. Yao C, Bai X, Liu W, Ma Y, Tu Z.'Detecting texts of arbitrary orientations in natural images', 2012 IEEE conference on computer vision and pattern recognition.2012; 1083–1090.
29. Khan T, Mollah AF.'A novel text localization scheme for camera captured document images', Proceedings of 2nd International Conference on Computer Vision & Image Processing.2018; 253–264.
30. Nikitin F, Dokholyan V, Zharikov I, Strijov V.'U-net based architectures for document text detection and binarization', International Symposium on Visual Computing.2019; 79–88.
31. Nagaoka Y, Miyazaki T, Sugaya Y, Omachi S. 'Text detection by faster R-CNN with multiple region proposal networks', 2017 14th IAPR international conference on document analysis and recognition (ICDAR).2017; 6, 15–20.
32. Risnumawan A, Shivakumara P, Chan CS, Tan CL.'A robust arbitrary text detection system for natural scene images', Expert Systems with Applications.2014; 41(18), 8027–8048.
33. Sun L, Huo Q, Jia W, Chen K. A robust approach for text detection from natural scene images. Pattern Recogn. 2015;48(9):2906–20.
34. Yi C, Tian Y.'Text detection in natural scene images by stroke gabor words', 2011 international conference on document analysis and recognition.2011;177–181.
35. Ma J. Arbitrary-oriented scene text detection via rotation proposals. IEEE Trans Multimedia. 2018;20(11):3111–22.
36. Cho H, Sung M, Jun B.'Canny text detector: Fast and robust scene text localization algorithm', Proceedings of the IEEE conference on computer vision and pattern recognition.2016; 3566–3573.
37. Zhu A, Gao R, Uchida S.'Could scene context be beneficial for scene text detection?', Pattern Recognition. 2016; 58, 204–215.
38. Sengar SS, Mukhopadhyay S. Motion segmentation-based surveillance video compression using adaptive particle swarm optimization. Neural Comput Appl. 2020;32(15):11443–57.
39. Prasad S, Kong AWK,'Using object information for spotting text', Proceedings of the European Conference on Computer Vision (ECCV).2018; 540–557.
40. Wu H, Zou B, Zhao Y-Q, Chen Z, Zhu C, Guo J.'Natural scene text detection by multi-scale adaptive color clustering and non-text filtering', Neurocomputing. 2016; 214, 1011–1025
41. Li H, Doermann D, Kia O.'Automatic text detection and tracking in digital video', IEEE transactions on image processing. 2000; 9(1), 147–156
42. Sharma N, Shivakumara P, Pal U, Blumenstein M, Tan CL. 'A new method for arbitrarily-oriented text detection in video', 2012 10th IAPR International Workshop on Document Analysis Systems, 2012, 74–78.
43. Sengar SS. 'Motion segmentation based on structure-texture decomposition and improved three frame differencing,' In IFIP International Conference on Artificial Intelligence Applications and Innovations, 609-622, 2019. Springer, Cham.
44. Carbonell M, Mas J, Villegas M, Fornés A, Lladós J. 'End-to-end handwritten text detection and transcription in full pages', 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW).2019; 5, 29–34.
45. Guha A, Samanta D.'Real-time application of document classification based on machine learning', International Conference on Information, Communication and Computing Technology.2019; 366–379.
46. Guha A, Samanta D, Banerjee A, Agarwal D. 'A deep learning model for Information Loss Prevention from multi-page digital documents', IEEE Access.2021.
47. Sengar SS, Hariharan U, Rajkumar K. 'Multimodal biometric authentication system using deep learning method,' In 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 309-312, 2020 IEEE.
48. Guha A, Samanta D.'Hybrid Approach to Document Anomaly Detection: An Application to Facilitate RPA in Title Insurance', International Journal of Automation and Computing. 2021;18(1), 55–72
49. Neumann L, Matas J.'Efficient scene text localization and recognition with local character refinement', 2015 13th International Conference on Document Analysis and Recognition (ICDAR).2015; 746–750.
50. Neumann L, Matas J. 'A method for text localization and recognition in real-world images', Asian conference on computer vision, 2010; 770–783.