International Journal of Computational and Experimental Science and Engineering (IJCESEN)

Vol. 11-No.3 (2025) pp. 4713-4721
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Enhanced Textual Data Reconstruction from Scanned Receipts Using Normalized Cross-Correlation and Deep Learning-Based Recognition with Superior Analytical Robustness and Computational Efficacy

M. Kathiravan[1*], A. Mohan[2], M Vijayakumar[3], M. Manikandan[4], Terrance Frederick Fernandez[5], Arumugam S S[6]

[1]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Thandalam, Chennai-602 105. India
* **Corresponding Author Email:** kathiravanm.sse@saveetha.com, **ORCID:** 0000-0002-5377-7871

[2]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Thandalam, Chennai-602 105. India
Email: annamalaimohan@gmail.com, **ORCID:** 0000-0003-1800-5371

[3]Department of Computing Technologies, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai – 603203
Email: vijayakm10@srmist.edu.in, **ORCID:** 0000-0003-1192-0331

[4]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Thandalam, Chennai-602 105. India
Email: manikandanm10@gmail.com, **ORCID:** 0000-0003-1081-0958

5Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Thandalam, Chennai-602 105. India
Email: frederick@ptuniv.edu.in, **ORCID:** 0000-0002-7317-3362

[6]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Thandalam, Chennai-602 105. India
Email: ssarumugam.me@gmail.com, **ORCID:** 0000-0002-0233-3832

Article Info:

DOI: 10.22399/ijcesen.3284
Received : 29 April 2025
Accepted : 01 July 2025

Keywords

Optical Character Recognition
Auto Text Extraction
Normalized Cross Correlation
Template Matching
Deep Learning

Abstract:

Text extraction from images plays a crucial role in optical character recognition applications such as invoices and receipt recognition. The recent character recognition approaches work well for good-quality scanned receipts, but they fail to do the same for low-quality receipts, offering reduced accuracy instead. This paper proposes invoice receipt identification using normalized cross-correlation-based template matching and a novel auto-text extraction approach using a deep learning algorithm. The proposed technique includes three major steps, preprocessing, character recognition and post-processing. The first step, which commences with preprocessing, involves noise removal, quality enhancement and image de-skewing. In the second step, auto-text extraction is carried out using a deep learning algorithm. The final post-processing step includes configuring the extracted text and exporting it to Word/Excel. According to the experimental results, the accuracy of the proposed approach outperformed existing approaches.

1. Introduction

Document digitization offers an invaluable means to develop and maintain data in the form of scanned documents [1]. Access to today's state-of-the-art technology has facilitated the use of massive volumes of scanned documents, significantly cutting the use of paper-based, printed documentation [2]. This is aggravated by problems in managing physical document storage, occasionally resulting in paper-based documents getting lost or damaged. Further, storing colossal volumes of such documents demands vast storage spaces [3].

Recognizing and extracting the required attributes from invoice receipts and making them available in the form of structured information is fundamental [4]. An automated framework using receipt information is employed for accounting or taxation

[5, 6]. Optical Character Recognition (OCR) has applications in a slew of domains when combined with deep learning. Using OCR to generate invoice receipts is still a work in progress, owing to the need for greater accuracy in extraction. The rapid and consistent implementation of OCR could steadily eliminate human jobs [7]. OCR also recognizes text characters from scanned documents [8] [9], such as invoices, bills, or receipts. Documents can be structured or semi-structured [10] [11], multi-format, pdf, jpg, or in an image file format. Extracting text using OCR is insufficient for powering, designing, or implementing data-driven machine learning models or operations. Information pertaining to, for instance, the date, payee name, total amount, and product list, among others, is to be extracted. Extracting primary parameters from documents plays a major role in services and applications like digitizing documents and converting the information in the documents into structured or semi-structured databases. The databases help in archiving, rapid indexing, and comparative analysis of the data [12] [13]. Abstract Information Extraction from Scanned Invoices (AIESI) is invaluable in dealing with document-intensive tasks in accounting, finance, law and medicine. There has been breakthrough research in OCR with respect to model accuracy and latency. AIESI helps automate data archiving and streamlines primary parameters from documents. For AIESI to be commercially viable, however, high accuracy and low latency are required to process large numbers of documents. Currently, data extraction from documents is manually processed - consequently, it is subject to bias and ambiguity in identifying key parameters. AIESI resolves the problems above with higher accuracy and lower ambiguity than humans are capable of in identifying key parameters. While manual work is time-consuming, AIESI processes millions of documents in a matter of hours and with the highest accuracy. Further, while human understanding is limited when dealing with multilingual documents, AIESI design models support such documents easily [14]. The proposed work uses a Convolutional Neural Network (CNN) based model for text extraction while preprocessing and OCR algorithms help extract the required information from the text. The results are obtained and a performance comparison is made using the Scanned Receipts OCR and Information Extraction (SROIE) dataset [15].

2. Literature Work

A Google Vision OCR-based information extraction approach is developed for large numbers of scanned documents. The extraction framework starts with loading input documents from Apache Hadoop, and the outcomes are saved in the Hadoop Distributed File System (HDFS) [16][17]. Google Vision OCR achieved 100% extraction accuracy in less time than other OCR tools, thus demonstrating the efficiency of its approach [18]. The advantage of deep learning in computer vision is that it identifies and extracts scanned invoice receipts, as discussed in this study. The recognition approach has two modules, text detection based on the Connectionist Text Proposal Network (CTPN), and text recognition based on the Attention-based Encoder-Decoder (AED). This method achieved an F1 score of 71.9% for the detection and recognition tasks [19].

To extract key information such as payee name, invoice number, total amount, items, address, and date from a document, AIESI eliminates the need for manual effort. It facilitates key attribute extraction from scanned documents with ease. Experimental results demonstrated a performance gain over other methods for key information extraction. A multimodal neural network is designed to learn from word embeddings obtained by the OCR from the image. This method maximized accuracy by 3% without clean text information. An innovative technique that scans invoice documents for attribute extraction is presented. It permits instantaneous textual encoding, layout and illustration information that is fed to the segmentation method. For the information extraction task, Visual Word Grid showed better results than state-of-the-art models on two (public and private document image) datasets [20].

Document Text Localization Generative Adversarial Nets (TLGAN) uses neural networks to localize text from receipts. TLGAN is an adaptable and flexi-train text localization method that needs little data. The approach attained 83% precision for the SROIE test data [2]. Prior to post-OCR parsing, a consolidated receipt data parsing technique is presented as a primary task. The dataset includes thousands of images of receipts and text annotations as well as their labels for parsing [6] [21]. A methodology is presented through which any information can be extricated from a printed invoice. The intermediate image is passed over using an OCR engine for further processing. The segmentation process extracts written text in various fonts and languages. Key data extraction from bills is carried out using Open CV and the Tesseract OCR engine [22] [23] [24] [25]. Similarly, While the articles [26] [27] [28] [29] have been proposed various techniques to extract text information from receipts, [30][31][32][33][34] were proposed to recognize the images from receipts for the purpose of information retrieval and classification of text images. Eventually, While the ensemble classifier [35]

extracts embedded knowledge from diversity of web sources, a new feature extraction method [36] [37] has been developed for classifying audio data.

3. Material and Methods

A knowledge extraction process from a dataset is carried out manually in the current framework. In defining key parameters, manual processing can be biased and ambiguous, which are problems that AIESI can resolve. The precision of the parameters considered is higher than that set by a person, with less uncertainty. Manual processing becomes overwhelming asthe passage of time passes, while automation processes millions of documents precisely in hours. While human comprehension falls short when dealing with multilingual documents, AIESI-designed models have no such problems. Data analysis may be used to derive a sound understanding of customer behaviour from such texts. The proposed text extraction block diagram for OCR scanned invoice receipts is shown in Fig. 1. The work comprises two stages, vendor identification and attributes extraction. Scanned structured or semi-structured invoice receipts are recognized, extracted and saved as structured data with uses across several applications. Data analysis may be used to derive a sound understanding of customer behavior from such texts. The invoice receipt information obtained can be used to automate office paperwork, including accounting and taxation. While tasks such as card and license plate recognition have improved markedly with recent advances in OCR, Receipt OCR remains challenging because of its increased accuracy requirements. In some instances, the recognition method also has problems with receipts of limited sizes and handwritten ones. Fig. 1 displays two small-sized and handwritten examples of receipts, with manual marking, still commonly used for these purposes.

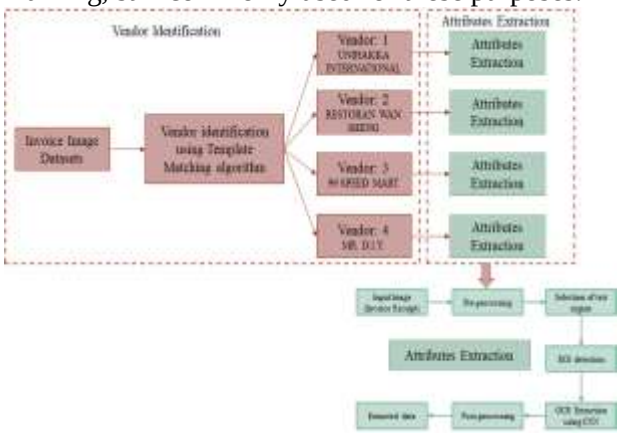

Figure 1: *Block diagram for automatic text extraction process*

3.1 Input Image

The International Conference on Document Analysis and Recognition (ICDAR) SROIE dataset original is a data package. Approximately four main text fields are incorporated in each image receipt, including the vendor's name, date and invoice number. The text in the invoice dataset consists primarily of numbers and English characters. Fig. 2 shows the invoice datasets of four different vendors.

Figure 2*: Invoice datasets*

3.2 Vendor Identification

The proposed method considered the invoice receipts of the following 4 vendors: Unihakka International, Restoring Wan Sheng, 99 Speed marts and Mr. D.I.Y. Vendor identification is undertaken using Normalized Cross-Correlation (NCC), the best template matching algorithm. Template matching detects the position of a template image inside an input image.

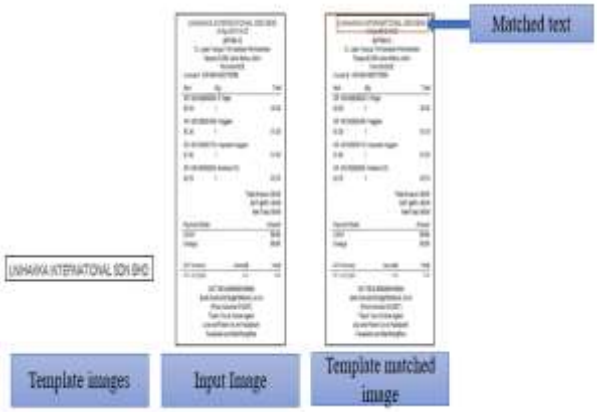

Figure 3: *Vendor identification using template matching*

Fig. 3 shows vendor identification using template matching. Cauchy-Schwartz's inequality makes the template matching algorithmic process

straightforward. The processing steps of the algorithm are listed as follows.

Step 1: Read the original and template images.

Step 2: Add zeros in all four corners of the image such that the template falls into the center of the original image.

Step2.1: Compute the template dimensions.

Step 2.2: Then, shift the template mask over the complete input image.

Step 2.3: Compute the square and sum of all the values of the padded image under the template.

Step 3: Now, move the mask over the entire image and concurrently compute the values of summation of template padded image under the template and store it in an array.

Step 4: Calculate the values of the padded image under the template's square and sum all values. Take the square root of the attained value and store it in an array.

Step 5: Compute the result value by dividing the result attained in (iii) by the result attained in step 4.

Step 6: Identify the region where the maximum value falls in the result above. The coordinates procure a perfect match of the template image and find the highest cross-correlation coefficient.

Step 7: Get the template images onto the main image using the coordinates and template size acquired above.

3.3 Preprocessing

These four major preprocessing measures include grayscale conversion, orientation correction, noise elimination and thresholding. The preprocessing flow is shown in Fig. 4.

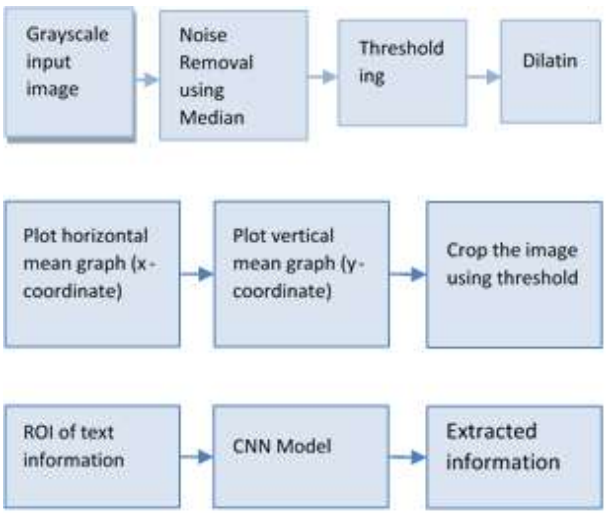

Figure 4:*The preprocessing steps*

3.4 Grayscale Conversions

Transforming a color image into a gray image involves a deep understanding of the former. The three colors in a picture are the red, green, and blue (RGB) pixel colors. The RGB color values are seen in XYZ in three dimensions, with the lightness, chroma and hue characteristics illustrated. The color image quality is based on the color reflected by the number of bits a digital unit can accommodate. A simple 8-bit, a high-color 16-bit, an actual 24-bit and a deep 32-bit color image are represented. The number of bits determines how many colors a digital interface allows. If each of the red, green and blue colors is 8 bits, the RGB combination takes 24 bits and supports 16,777,216 colors. These 24-bit constitutes the luminance in a color picture of a pixel, while the grayscale displays no more than 8 bits of luminance. The luminance ranges from 0 to 255 for a pixel value of a gray image. RGB (24-bit) values are converted into (8-bit) grayscales when a color picture is transformed into a grayscale. Different methods are used to convert an RGB image into a grayscale image, such as, for instance, the average method and the weighted method.

3.4.1 Average Method

The average value of the system is the R, G, and B average. Grayscale = (R + G + B) / 3. Theoretically, the formulation is 100% accurate. When writing code, however, the overflow error of unit 8 is found as the sum of R, G, and B that is more than 255. Grayscale = R / 3 + G / 3 + B / 3. The average approach is straightforward but does not work as expected, owing to the different reactions the human eye has to RGB hues. The eyes are most sensitive to green, less so to red, and least of all to blue.

3.4.2 Weighted Method

The weighted form, also called the method of luminosity, weighs red, green, and blue wavelengths. The formula used is the following: Grayscale = 0.299R + 0.587G + 0.114B.

3.5 Orientation Correction

The orientation of a page is determined by the direction in which paper is put in the scanner. During scanning, a document can be fed to the scanner in a particular direction by positioning it accordingly. The document feeding in different directions results in the image being appropriately oriented. Text that is not in the correct orientation, i.e., except 00-page orientation, is oriented text. Text on a 900 or 1800 or 2700 page may be focused. Feeding the paper in the direction of the portrait introduces a near 00 or 1800 image orientation. Similarly, landscape feeding introduces a near 900 or 2700 orientation.

Documentation text-oriented reading or extraction adversely impacts existing OCR systems. A considerable body of research has been published on the skew and orientation detection. Detecting page orientation is easy if the layout of the paper is known. This information is typically inaccessible, and page orientation needs to be extracted from the documents themselves, especially from global features. Researchers have tried to detect the direction of the page, with each approach having its constraints and benefits. The methods all contribute to finding the skew angle and propose rotating the image of the document in the opposite direction of the observed angle of inclination. The technique rotates an image by a certain angle and a given axis or point, as is plain from its name. It converts a geometrical position of a point in the current image by rotating it through the user angle of the defined axis to the output image. The points beyond the output picture boundary are ignored, and rotation is chiefly used to improve visual appearance.

3.6 Noise Removal

Spatial frequency is a distance sensor function, expressed as a number for a specific portion of an image whose brightness value changes per unit of distance. A given region in an image that is scarcely altered by the brightness value is known as a low-frequency area. Conversely, the brightness value that varies greatly over short distances is a high-frequency range. Spatial filtration separates the image into spatial frequencies and selectively adjusts certain frequencies to underline specific image characteristics. This method improves the analyzer's capacity to discriminate between information. The nonlinear median filter used here is simple and efficient, with the following advantages:

It minimizes the difference in intensity between one pixel and another.

- The value of the median pixel is substituted.
- The mean is determined when the entire pixel values are sorted and replaced by the middle pixel value with the pixel value measured.

The mean filtering process is quick, intuitive and easy to use, i.e., it minimizes the difference in intensity between one pixel and the next. It is typically used to reduce image noise.

3.7 OCR Extraction Using the CNN

A CNN is usually put to work on the image data, with each image made up of a matrix of pixels. Depending on the bit rate, the number of values to be encoded for each pixel is considered. The possible range of values expressed by a single pixel is, therefore, [0, 255]. The existence of separate colour channels (3 for RGB images), particularly RGB (red, green, blue) images, however, introduces a further 'depth' field to the data, which makes the entry three-dimensional. The picture thus constitutes an entire 3D structure called the quantity of the input image (255x255x3).

- The convolution layer, which forms the base of the CNN, conducts the key training process and consists of neurons organized in different layers. As defined in the previous section, it performs a complicated operation over the input quantity and consists of a 3D neuron collection.
- Each neuron network is linked to a positive state called the receptive field of the input volume. For example, in a 28x28x3 input image, if the field of interest is 5x5, then every neuron in the convolutional layer is linked to a 5x5x3 state in the input volume. Each neuron would, therefore, have 75 weighted inputs.
- There is a section of neurons opposite that is entirely devoted to capturing input from this state for a specific R value.

3.7.1 The Rectified Linear Unit layer

ReLu defines the rectifier unit, which is the activation mechanism most widely used. For CNN neuron outcomes, and defined mathematically as:

$$\text{Max}(0, x) \quad (1)$$

The ReLu, at the point of its origin, is not distinguishable, which makes it difficult to be used for back propagation training.

$$f(x) = \ln(1+e^{x}) \quad (2)$$

As stated in a prior blog post, the derivative of the soft plus feature is the sigmoid function.

$$f(x) = (d(\ln(1+ e^{x})))/dx = e^{x}/(1+e^{x}) = 1/(1+ e^{x}) \quad (3)$$

3.7.2 The Pooling Layer

The pooling layer normally follows the convolution layer. Its usefulness lies in reducing the image dimensions of the input volume for the convolution layer, though the depth dimension of the volume is unaffected as a result. The process performed by this layer is often referred to as down sampling, and the size reduction contributes to information loss. Such a loss is, however, advantageous to the other

network layers, where the diminished size results in reduced overhead computing and avoids overfitting. The pooling layer makes a sliding window that is pushed in step across the input, translating the values into representative results, much like the convolution procedure performed above. The conversion is achieved either by taking the highest value out of the values that can be found in the window or by taking the average of the values. Owing to its superior performance features, max pooling has been favoured over others. The operation is executed for each depth slice. For instance, if the input size is 4x4x3 and the sliding window size is 2x2, the max pooling operation performed for each colour channel down samples the values to their representative highest value. The operation employs a function over the input values, taking fixed portions at a time, and is adaptable as parameters with a scale. In CNNs, grouping is optional, and much of the architecture does not perform pooling operations.In Fig. 5, sub-figures (ii) and (iii), the max pooling process can be observed to max-pool the 3 colour channels to indicate the input volume for the pooling layer. The procedure uses a [2, 2] stride value. The window motion is defined by the dark and red boundary regions. The operation applied for a stride value of [1, 1] is shown in sub-figure 5, resulting in a matrix of 3 to 3.

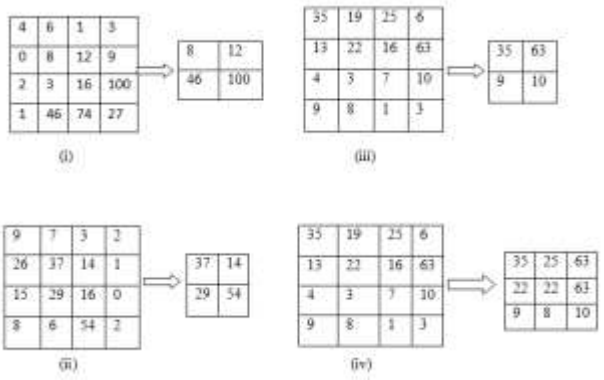

Figure 5: *The operation applied*

3.7.3 The Fully Connected Layer

The completely connected layer, which is precisely what its name implies it is, is fully linked to the results of the prior layer. Usually, fully connected layers connect the output layer and generate the preferred number of outputs in the last stages of the CNN.

3.8 Post Processing

String matching is primarily applied to process text for various applications. In such a fundamental process, major issues arise when identifying identical DNA sequences. Given a text T [1, . . .,n]

String Matching Algorithm (P[1, . . . , m], T[1, . . .,n])

Input: pattern P of length m and text T of length n

Preconditions: $1 \le m \le n$

Output: list of all numbers s, such that P occurs with shift s in T

for s ← 0 to $n - m$

{

if (P[1, . . . , m] == T[s + 1, . . . , s + m])

{

Output s

}}

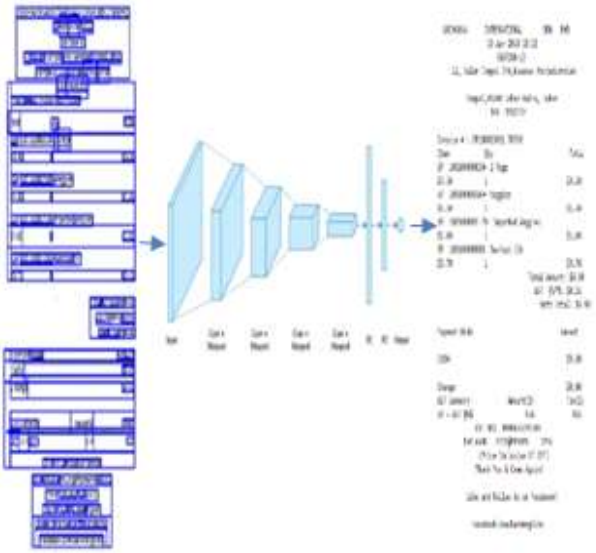

Figure 6: *Text extraction using OCR*

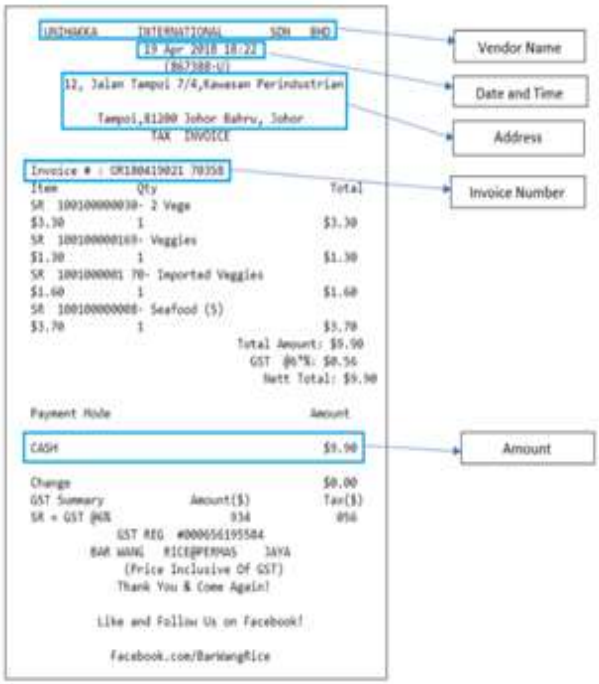

Figure 7: *Information extraction from image*

Table 1: *Information Extraction*

S.No	Vendor Name	Date	Address	Invoice Number	Amount
1	Unihakka International	19-04-2018	12, Jalan Tampoi, 7/4, Kawasan Perindustrain, Tampoi, 81200 Johor Bahru, Johor	OR180419021 7038	9.90
2	Restoran Wan Sheng	30-08-2018	No.2 Jalam Temenggung 19,9, Seksyen 9, Bandar Mahkota Cheras, 43200, Cheras,Selangor	12194461	RM:4.20
3	99 Speed Mart	07-02-2017	Lot P.T. 2811, Japan Angsa, Taman Berkeley, 41150 Klang, Selangor	17936/102/T0275	RM:37.4
4	MR.D.I.Y SDN BHD	28-05-2018	Lot 1851-A, 1851-B, Jalankpb 6, Kawasan Perindustrian Balakong, 43300, Serikembangan, Selangor	R000059251	RM: 2.70

and a pattern $P[1, \ldots, m]$, the entire possibilities of P are computed in T. P arises in T with shift s if $P[1, \ldots, m] = T[s + 1, \ldots, s + m]$. All possible comparisons are made by the equation $(n - m + 1)\, m = \Theta(nm)$. For the string matching issue, each symbol in T and P must be observed in $\Omega(n + m)$ time. While Fig. 6 represents the text extraction process using OCR, Fig. 7 shows the identification of attributes from the image taken for extraction. The extracted information attributes are represented in Table 1.

4. Results and Discussions

The results show that the proposed model outstrips the existing system for extracting key parameters from scanned documents. Given the assembled visual and spatial features, this approach will mark each bounding box with better precision. The exactness of the bounding box and the text depend on the accuracy of the entire pipeline. Accuracy remains low if the OCR is below par, regardless of the accuracy of our model. Here, the SROIE dataset is used and the bounding box extracted. It is hard to eliminate the bounding box with 100% accuracy in a real-life environment; However, Table 2 and Fig. 8 show the extracted results and a comparison of accuracy, respectively.

Table 2: *Extracted Results*

Vendor Name	UNIHAKKA INTERNATIONALSDN BHD
Date and Time	19-04-2018 18:22
Address	12, Jalan Tampoi 7/4, Kawasan Perindustrian, Tampoi,81200 Johor Bahru, Johor
Invoice number	OR180419021 70358
Amount	$9.90

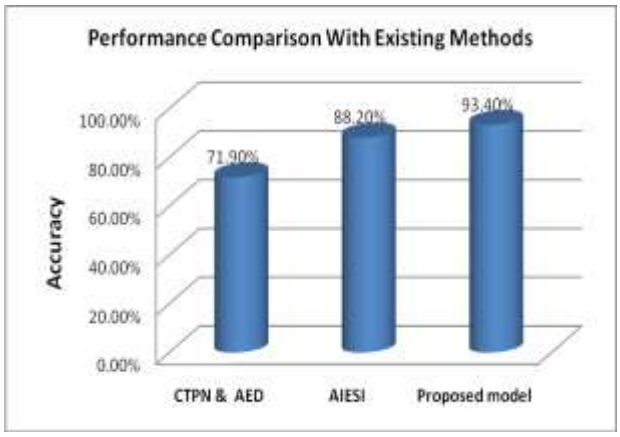

Figure. 8: *Performance comparison with the state-of-the-art approaches*

In Fig. 8, while the existing approaches secured 71.9 and 88.2 accuracies, the proposed model has secured 93.4%, which is 21.3 % higher than CTPN for text detection & AED for text recognition and 5.2 % larger than abstractive information extraction from scanned invoices. Though many existing algorithms give good results, our model has achieved a higher result than other models.

5. Conclusion

This research has explored a limited number of approaches to certain problems faced by OCR systems in automated document reading. The focus has been on preprocessing images for input documents in order to prepare them to enhance the readability of OCR systems. The literature reviewed has briefly described the research carried out in this direction. Preprocessing and OCR were critical to

processing document files. A thorough analysis has been undertaken on a few available OCR systems to demonstrate their limitations. Given that OCR systems read horizontally linear text, it is necessary that text in documents be so prepared before it can be read. Current research aims to make OCR systems more effective by preprocessing input documents. Further, we studied conversion models that make interpretation by OCR systems easier when applied to document input images. The research was also intended to resolve the shortcomings/vulnerabilities of OCR systems, resulting in incorrect transcriptions of input images. The research was also intended to resolve the shortcomings/vulnerabilities of OCR systems, resulting in incorrect transcriptions of input images. Moreover, the overall efficiency of OCR systems can be enhanced by adding amendments. This study has thrown open new perspectives for analysis. There is scope for great improvement, particularly FOR designing an improved filter for noise reduction and enhanced contrast and including 5 separate seller bills (seller's name, date and time, invoice number, address and sum) in the list.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Automatic receipt recognition system based on artificial intelligence technology. *Applied Sciences, 12*, 853. https://doi.org/10.3390/app12020853

[2] Patel, S., & Bhatt, D. (2020). Abstractive information extraction from scanned invoices (AIESI) using end-to-end sequential approach. *arXiv preprint arXiv:2009.05728*. https://arxiv.org/abs/2009.05728

[3] Le, A. D., Van Pham, D., & Nguyen, T. A. (2019). Deep learning approach for receipt recognition. In *International Conference on Future Data and Security Engineering* (pp. 705–712). Springer. https://doi.org/10.1007/978-3-030-35649-1_48

[4] Audebert, N., Herold, C., Slimani, K., & Vidal, C. (2019). Multimodal deep networks for text and image-based document classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 427–443). Springer. https://doi.org/10.1007/978-3-030-43887-6_27

[5] Kerroumi, M., Sayem, O., & Shabou, A. (2020). Visual word grid: Information extraction from scanned documents using a multimodal approach. *arXiv preprint arXiv:2010.02358*. https://arxiv.org/abs/2010.02358

[6] Kim, D., Kwak, M., Won, E., Shin, S., & Nam, J. (2020). TLGAN: Document text localization using generative adversarial nets. *arXiv preprint arXiv:2010.11547*. https://arxiv.org/abs/2010.11547

[7] Park, S., Shin, S., Lee, B., Lee, J., Surh, J., Seo, M., & Lee, H. (2020). CORD: A consolidated receipt dataset for post-OCR parsing. *arXiv preprint arXiv:2005.00642v3*.https://arxiv.org/abs/2005.00642

[8] Sharma, S., Gaur, M., Kumar, Y., & Varma, M. (2022). A hybrid model for invoice document processing using NLP and deep learning. *International Journal of Computer Applications, 184*(19), 9–14. https://doi.org/10.5120/ijca2022922440

[9] Majumder, A., Singla, A., & Mahajan, A. (2021). Structured information extraction from scanned documents using deep learning. *Procedia Computer Science, 192*, 3403–3412. https://doi.org/10.1016/j.procs.2021.09.111

[10] Burie, J. C., Nicolaou, A., Rusiñol, M., Karatzas, D., Mouchère, H., & Vincent, N. (2017). ICDAR2017 competition on recognition of documents with complex layouts—RDCL2017. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 1403–1410). IEEE. https://doi.org/10.1109/ICDAR.2017.231

[11] Huang, Z., Liu, W., Li, J., Li, Z., & Li, H. (2022). Document information extraction using BERT with layout features. *Pattern Recognition, 128*, 108676. https://doi.org/10.1016/j.patcog.2022.108676

[12] Katti, A. R., Reisswig, C., Guder, C., Brarda, S., Höhne, J., Bickel, S., & Faddoul, J. B. (2018). Chargrid: Towards understanding 2D documents. *arXiv preprint arXiv:1809.08799*. https://arxiv.org/abs/1809.08799

[13] Xu, Y., Xu, J., Lv, T., Cui, L., Wei, F., Wang, Y., ... & Zhou, M. (2020). LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1192–1200). https://doi.org/10.1145/3394486.3403172

[14] Xu, Y., Lv, T., Cui, L., Lu, Y., Lu, Y., Wei, F., & Zhou, M. (2021). LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (pp. 2579–2591). https://doi.org/10.18653/v1/2021.acl-long.203

[15] Xu, Y., Lv, T., Cui, L., Lu, Y., Wang, G., & Zhou, M. (2022). LayoutXLM: Multimodal pre-training for multilingual visually-rich document understanding. *Findings of the Association for Computational Linguistics: ACL 2022*, 1632–1644. https://doi.org/10.18653/v1/2022.findings-acl.130

[16] Garncarek, Ł., Powalski, R., Stanisławek, T., & Śmieja, M. (2021). LamBERT: Layout-aware

language modeling for information extraction. *arXiv preprint arXiv:2002.08087*. https://arxiv.org/abs/2002.08087

[17] Powalski, R., Stanisławek, T., Grabowski, P., & Garncarek, Ł. (2021). Donut: Document understanding transformer without OCR. *arXiv preprint arXiv:2111.15664*. https://arxiv.org/abs/2111.15664

[18] Appalaraju, S., & Chao, C. (2021). DocTr: Document image transformer for geometric unwarping and text recognition. *arXiv preprint arXiv:2106.03060*. https://arxiv.org/abs/2106.03060

[19] Hong, J., Lee, J., Lee, S., Yim, J., & Park, S. (2022). Bros: A pre-trained language model focusing on text and layout for better key information extraction from documents. *Pattern Recognition, 129*, 108766. https://doi.org/10.1016/j.patcog.2022.108766

[20] Lee, B., Yim, J., Park, S., Kim, G., Shin, J., Lee, S., & Hong, J. (2022). LiteBERT: An efficient pre-trained language model for document understanding. *arXiv preprint arXiv:2202.13634*. https://arxiv.org/abs/2202.13634

[21] Hwang, W., Yim, J., & Park, S. (2021). Spatial-aware BERT for document-level layout understanding. *arXiv preprint arXiv:2103.14470*. https://arxiv.org/abs/2103.14470

[22] Lee, B., Yim, J., & Park, S. (2021). PILOT: Position Aware Text Representation for Key Information Extraction. *arXiv preprint arXiv:2103.12213*. https://arxiv.org/abs/2103.12213

[23] Ahmad, W., Chakraborty, T., Yuan, X., & Chang, K.-W. (2019). Context-aware layout analysis for document image understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 0–0). https://doi.org/10.1109/CVPRW.2019.00113

[24] Khan, M. A., Akram, T., Zhang, Y.-D., & Sharif, M. (2021). Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework. *Pattern Recognition Letters, 143*, 58–66. https://doi.org/10.1016/j.patrec.2020.12.015

[25] Ramshaw, L. A., & Marcus, M. P. (1995). Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora* (pp. 82–94). https://aclanthology.org/W95-0107/

[26] Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 260–270). https://doi.org/10.18653/v1/N16-1030.

[27] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186). https://doi.org/10.18653/v1/N19-1423

[28] Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*. https://arxiv.org/abs/1606.05250

[29] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Dollár, P. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

[30] Afzal, M. Z., Kölsch, A., Ahmed, S., & Dengel, A. (2017). Cutting the error by half: Investigation of very deep CNN and advanced training strategies for document image classification. In *14th IAPR International Conference on Document Analysis and Recognition (ICDAR)* (Vol. 1, pp. 883–888). IEEE. https://doi.org/10.1109/ICDAR.2017.148

[31] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2961–2969). https://doi.org/10.1109/ICCV.2017.322

[32] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*. https://arxiv.org/abs/1804.02767

[33] Zhang, Y., Qiu, M., Chen, Y., & Huang, J. (2018). End-to-end information extraction based on deep reinforcement learning. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics* (pp. 1–12). https://aclanthology.org/C18-1001/

[34] Liu, P., Yuan, H., Fu, J., Jiang, Z., & Zhang, Y. (2021). Structured information extraction from noisy semi-structured documents using structure-aware pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6177–6186). https://doi.org/10.18653/v1/2021.emnlp-main.500

[35] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. https://arxiv.org/abs/1907.11692

[36] Su, J., Xu, Y., Li, S., Cui, L., Wang, G., Wei, F., & Zhou, M. (2021). VIES: A novel video information extraction system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 4375–4383). https://doi.org/10.1145/3447548.3467323

[37] Zhang, Y., Xu, J., & Cui, L. (2020). LayoutLM: A unified model for understanding documents. In *arXiv preprint arXiv:2004.14797*. https://arxiv.org/abs/2004.14797