# A study on automated invoice recognition and text correction

Jin-Se Lee[1] · Kyu-Wan Han[2] · Hong-Il Seo[3] · Dong-Hoan Seo[†]

**Abstract:** Recently, with the growth of the domestic parcel delivery and e-commerce markets, there has been an increasing demand for automated parcel and package acceptance systems. The post office and convenience stores are using kiosks to replace them, but the usage rate of middle-aged and elderly people, who are the main users who are not familiar with digital devices, is low, making it difficult to call it a reasonable automation method. In addition, the existing paper-based receipt system is insufficient to meet modern requirements. Therefore, this paper proposes a system that automatically corrects and digitizes handwritten invoices and package information. The proposed system extracts invoice images that are resistant to optical and physical noise through an image segmentation model, and uses a visual document understanding model to correct address information that has been incorrectly recognized through an LLM-based correction algorithm to output the address. In particular, the system presents strategies for improving and commercializing the overall system by extracting areas of the invoice that were previously inferred using image processing algorithms alone, and by pre- and post-processing display models for handwriting recognition. This improves the performance of systems that process handwritten invoices and enables the detection and correction of errors and typos.

**Keywords:** Handwritten, Invoice recognition system, Segmentation, Visual document understanding, Large language model

## 1. Introduction

The recent exponential growth of the domestic parcel delivery and e-commerce markets has resulted in a surge in demand for unmanned and automated parcel and package acceptance systems, which are essential for maintaining the quality of service in the logistics system. The majority of online services within the domain of logistics systems are automated in a manner that enables users to convey the information they require via chatbots that utilize Large Language Models (LLM). In the context of offline services, post offices and grocery stores are supplanting the functions of printing invoices, calculating delivery charges, and processing payments, which were previously performed by humans, with the use of kiosks. Nevertheless, full automation remains challenging due to the low usage rate of middle-aged and elderly individuals who are less familiar with digital devices. To address this issue, a system that enables users to oversee the entire process while aligning with the existing parcel delivery system is essential. Consequently, to automate the postal and parcel delivery system, a technology capable of recognizing and digitizing existing handwritten parcel bills from images captured by a camera is required.

Two distinct technologies are employed to facilitate the digitization of the information present on the bill. The first of these is image segmentation, which is responsible for detecting both the load and the bill within the image. The second is optical character recognition (OCR), which is tasked with recognizing the characters present on the bill. Image segmentation is a method of delineating the boundary of the object of interest from the background and surrounding noise. Image segmentation is employed to differentiate between parcels and invoices, thereby enhancing the precision of text recognition in invoices. With the advent of deep learning, image segmentation has demonstrated remarkable

† Corresponding Author (ORCID: http://orcid.org/0000-0003-3610-0356): Professor, Division of Electronics & Electrical Information Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, 727, Taejong-ro, Yeongdo-gu, Busan 49112, Korea, E-mail: dhseo@kmou.ac.kr, Tel: 051-410-4412

1 M. S. Student, Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, National Korea Maritime & Ocean University, E-mail: ljs171181@g.kmou.ac.kr, Tel: +82-51-410-4822

2 B. S. Student, Division of Electronics & Electrical Information Engineering, National Korea Maritime & Ocean University, E-mail: hantwin4@naver.com, Tel: +82-51-410-4822

3 Ph. D. Candidate, Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, National Korea Maritime & Ocean University, E-mail: seoluck77@gmail.com, Tel: +82-51-410-4822

Jin-Se Lee ᆞ Kyu-Wan Han ᆞ Hong-Il Seo ᆞ Dong-Hoan Seo

capability for separation through the utilisation of diverse neural networks, including Mask R-CNN and Graph Convolutional Network (GCN) [1][2]. Nevertheless, it remains challenging to entirely eliminate optical noise, such as the user's hands, feet, and shadows. Consequently, the model necessitates supplementation.

Optical character recognition (OCR) can be classified into two main categories: online and offline systems. Online systems employ supplementary data, including handwriting, writing order, and speed, to facilitate the recognition of handwritten text. Conversely, offline systems present a more challenging domain due to their exclusive acceptance of images with written text as input [3]. Recently, there has been a notable advancement in this field due to the integration of neural networks, which have demonstrated remarkable efficacy as a classification algorithm. However, OCR is currently limited to recognizing a single text, necessitating the enhancement of the model to align with the specific requirements of industrial application.

The objective of visual document understanding (VDU) is to identify text within documents that have been stored as images and to comprehend the documents themselves [4]-[7]. This method is currently being studied in a number of different ways, including the classification of document types, the labelling of sequences, and the classification of attribute values associated with each text, depending on the specific application in question. A VDU comprises a vision transformer and a language transformer [8][9], as it must learn both the structure of text and documents, in contrast to OCR models. The vision transformer extracts visual features from document images, while the language transformer is trained using these features to recognize text and output information about the document for additional purposes. This technology represents a significant innovation in the digitization of documents recorded in the past.

The objective of this research is to develop a handwritten invoice recognition system that enhances the accuracy of invoice recognition through the implementation of image segmentation and address correction. The proposed system employs image segmentation to identify the boundaries of the invoice. It then extracts the address, phone number, and name information of the sender and receiver from the invoice and corrects the address information. The image processing algorithm and YOLACT model [10] are employed for the purpose of image segmentation. Similarly, the borderline detection algorithm and VDU model are utilized for the extraction of address information. In the address correction process, FAISS is employed to ascertain the degree of

similarity, while the LLM model is utilized to rectify the address information in accordance with the prevailing address system. The LLM model invokes OpenAI's ChatOpenAI API to calibrate addresses in accordance with a specified template. This template delineates the characteristics of the current addressing scheme and requests that the given address be modified to align with the scheme.

The major contributions of this work are summarized as follows:

1. In this paper, we propose a handwritten invoice recognition system that improves invoice recognition performance through segmentation and enables address correction.
2. Extract the boundaries of invoices through segmentation and resize them to a uniform size to improve invoice recognition performance.
3. Improve VDU performance by correcting typos or incorrectly detected words through address correction based on matching with the real address DB.

## 2. Related Work

### 2.1 Handwritten Invoice Recognition

Handwriting recognition is difficult to achieve high accuracy due to the wide variation in handwriting styles of writers [11]. In particular, Korean has a higher variation in handwriting compared to English, consisting of 51 consonant and vowel combinations [12]. There are two methods for handwriting recognition: OCR, which converts syllable-by-syllable images into letters, and VDU, which recognizes the structure and content of a document by repeatedly utilizing it [5] [6] [13]-[15]. Compared to OCR, VDU can consider the class of each text and the structure of the entire document, allowing users to extract the information they need, thus improving the efficiency of document processing [4] [8] [16]. In particular, handwritten invoice recognition is noteworthy because it consists of addresses, names, and phone numbers, allowing for individual text recognition and classification of the text. This structural characteristic is advantageous for increasing the performance of handwriting recognition systems, which can effectively classify and analyze the details.

### 2.2 Address Correction Algorithm

Detecting errors for proper nouns such as names and phone numbers in invoices is a very difficult problem [17]. Addresses, on the other hand, follow a legally prescribed way of writing, so

even typos and partial words can be automatically corrected to find the correct address [18] [19]. South Korea's addressing system has two types of addresses: street addresses and street names, both of which consist of two to five words, excluding details such as the number. Since each word consists of the sum of a proper name and a municipality name, OCR errors can be corrected by analyzing the similarity of the address to a database [20]. Therefore, database matching-based correction and Large Language Model (LLM)-based methods can be applied [21] [22]. Database matching-based correction separates the text generated by OCR into words and searches for similar addresses in the database using each word as a keyword [23]-[25]. Candidate addresses are then derived by finding the intersection of the search results for each word, and further language model-based similarity analysis is performed to determine the final address. However, this method performs a phoneme-by-phoneme comparison of keywords, which can lead to large errors for a given address. LLM-based methods use a language model trained on an address database to estimate the final address according to the address system by prompting the language model with an address system and then typing the text generated by OCR [26] [27]. This method allows the model to understand different address formats and extract key information from the input text to accurately estimate the final address. This makes it robust to address word counts, missing addresses, and typos.

# 3. Proposed Work/Method

## 3.1 Overall Handwritten Invoice System

The proposed system aims to digitize the information about the sender and recipient in the handwritten invoice image, compare the information with the current address system, and output the enhanced address information. **Figure 1** shows the structure of the handwritten invoice system, which consists of five steps. First, in the box image step, a camera is used to collect the image of a delivery box with an invoice attached. In the second step, an image processing algorithm is used to detect the invoice attached to the box and extract the sender and receiver images in the invoice. The Sender-Receiver Information Extraction step digitizes the handwritten information and outputs the address, name, and
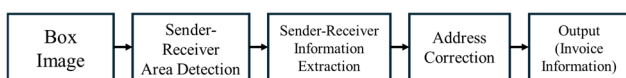
phone number using a neural network-based VDU model. In the Address Correction step, the address information output in the previous step is compared and enhanced with the current address system using FAISS and LLM models. At the end, the sender and receiver information including the corrected address information is provided to the user.

## 3.2 Box Image

The environment in which box images are collected is as shown in **Figure 2**. To detect invoices and recognize the text within them, it is necessary to maintain a minimum resolution. Therefore, we defined the minimum unit required to recognize a single character and used a camera capable of satisfying this requirement at a maximum distance of 60 cm between the camera and the invoice. **Figure 3** illustrates a sample of the images actually captured. To minimize environmental factors affecting box detection, images were collected in various locations such as on desks and floors. These captured images are then fed into the next step for invoice region detection.

## 3.3 Sender-Receiver Area Detection

In this step, we aim to detect the image of the region containing the address, name, and phone number of the sender and receiver using the box image. The proposed method consists of three modules, as shown in **Figure 4**. (Box image detection with information) module detects objects using the Yolact model. Each



**Figure 2 :** Invoice image collection environment
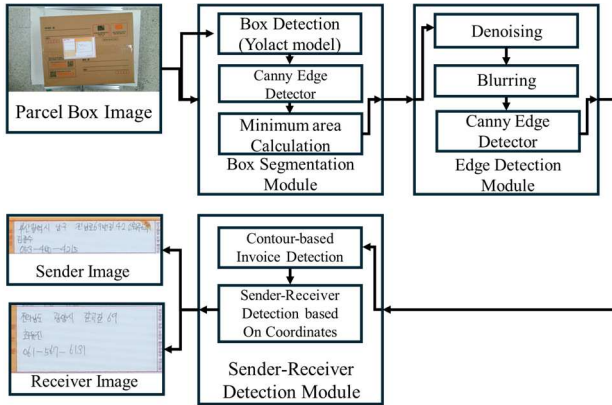


**Figure 3 :** : Sample image



**Figure 1 :** Overview of proposed system

Jin-Se Lee ・ Kyu-Wan Han ・ Hong-Il Seo ・ Dong-Hoan Seo



**Figure 4** : Sender-receiver image extraction process



**Figure 5 :** Sender-receiver information extraction process



**Figure 6 :** Address correction process

which performs correction on the JSON data output from the textual decoder.

### 3.5 Address Correction

The address correction module, as shown in **Figure 6**, consists of two stages: Similarity Matching and Address Correction. In the Similarity Matching stage, the input address is vectorized to calculate its similarity with a pre-constructed database. Based on this similarity, the top 5 most similar addresses along with their contextual information are extracted. This similarity calculation is efficiently performed using FAISS (Facebook AI Similarity Search). In the Address Correction stage, an LLM (GPT-4) is employed to correct errors in the input address using the extracted similar addresses and contextual information. During this process, the LLM distinguishes the structure of the Korean address system and corrects typos and formatting errors by reflecting contextual similarity. The output is generated based on a prompt explicitly specifying the structure of the Korean address system.

## 4. Experiment

We trained the Image Segmentation Model for sender-receiver region extraction and the VDU Model for digitizing sender-receiver information from handwritten invoices on a total of 16,000 train data of simultaneous courier and invoice images. The handwriting used in the invoices was generated based on address data created using a generative model trained on real handwriting. The test set consists of generated addresses as well as some cases of manually written address characters. In the case of the Image Segmentation Model, we checked its own performance through the Confusion Matrix, and in the case of the VDU Model, we compared its performance with and without the address correction algorithm proposed in the paper.

### 4.1 Quantitative Evaluation of Box Accuracy

We want to check the performance of the model by checking the Confusion Matrix for 3 classes: Box, Invoice, and Physical Noise for a total of 1,000 test data. The data was collected from

object is detected and edge images are detected using Canny Edge detector. Detect the minimum box area containing the invoice per bar by calculating the detected edge image and detect the minimum size box image containing the invoice. Then, the Edge Detection Module uses Gaussian blur to remove noise and blur using median filter to detect the edge of the invoice. Finally, the Canny Edge detector is used to detect the extracted edge image. The edge image and the denoised box image are input to the sender-receiver detection module. The module uses the edge image and box image to detect the invoice region and crops the region where the information about the sender and receiver is written in the invoice. The edge image is used to detect the coordinates of the four vertices of the rectangular invoice through contour extraction. Use this to extract the invoice region in the box image. Taking advantage of the fact that invoices are structured, crop the image containing the sender and receiver information. Both images are fed into the invoice understanding module.

### 3.4 Sender-Receiver Information Extraction

Sender-receiver information extraction is the process of outputting the information of the sender and receiver in the extracted invoice image. For information extraction, we used Donut, which performs well without requiring text location labeling. To reflect the domain of invoices, our model was trained on pre-trained Donut with additional created and collected data.

As shown in **Figure 5**, Donut consists of three main steps. When a Sender or Receiver image is input, it goes through the vision encoder step. The vision encoder step passes the Korean courier invoice through the encoder to extract visual features that contain detailed information such as the type and location of text in the image. The second step is the textual decoder, which is a language model that generates sentences with the output of the vision encoder and the prompt as input. The last step is postprocessing,
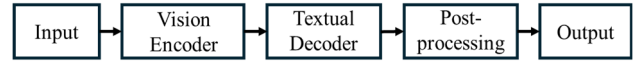
a camera that shoots perpendicular to the floor from a height of 60cm, and a UHD (3,840 * 2,160 pixel) camera was used to ensure that the letters on the invoice can be seen considering the height of the floor and the box. The boxes and invoices used in the experiments were parcel boxes numbered 1-4 and invoice paper based on the postal service headquarters. Image

Segmentation was performed by fine-tuning the train data based on YOLACT [10]. We checked the performance of the model by measuring the accuracy in mAP(50) and mAP(75) environments. mAP(50) and mAP(75) are metrics that evaluate model performance based on the Intersection over Union (IoU) value, which measures the overlap between predicted and ground truth boxes. Specifically, mAP(50) considers a prediction correct when the IoU is 0.5, while mAP(75) uses an IoU threshold of 0.75. These metrics help in understanding both the overall performance and precision of a model, with higher values indicating better performance. The results of the Baseline and the model with the proposed method are presented in **Table 1**.

**Table 1 :** mAP(50) and mAP(75) performance of Bounding Box and Segmentation Mask of Segmentation model

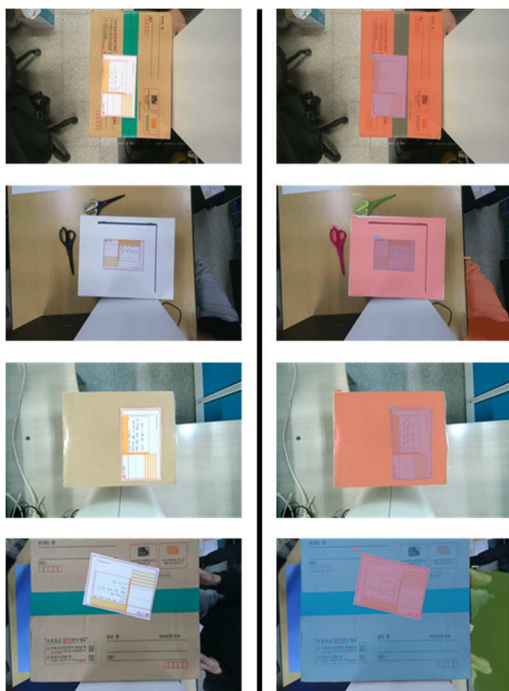| Type | mAP(50) (%) | mAP(75) (%) |
|---|---|---|
| Bounding Box | 92.47 | 87.78 |
| Segmentation Mask | 91.05 | 87.02 |



**Figure 7**: Image of delivery invoice, (left) original, (right) after segmentation

**Table 2 :** Number of correct and incorrect words & address accuracy of Baseline (Donut) model and Baseline + Address Correction (Ours)

| Model | Address Correct | Address Error | Address Accuracy (%) |
|---|---|---|---|
| Baseline [28] | 291 | 121 | 70.67 |
| Baseline + Ours | 312 | 100 | 75.72 |

## 4.2 Quantitative Evaluation of the Address Correction

To compare the performance of the resulting address correction algorithm to the existing Donut-based OCR, we measure the accuracy of each address in words. The data used for the test is about 100 images of the Sender-Receiver area that have been processed by the segmentation model. Accuracy is measured by comparing the model's predicted results with the actual data (ground truth, GT), which takes into account whether each character is matched or misspelled.

## 4.3 Qualitative Evaluation

For qualitative evaluation, we ran experiments on three bad cases: missing addresses, address typos, and handwriting. The initial predictions of the model and the address correction algorithm were as follows. Missing addresses refer to cases where specific elements of an address, such as a district or city, are missing or incorrect. Address typos refer to cases where a nonexistent address is written. In both cases, traditional OCR systems can only recognize the text as written, inevitably leading to address errors. Handwriting refers to cases where the address is overwritten, physically corrected, or written with unclear handwriting. In such scenarios, OCR systems frequently encounter errors, whereas our model is designed to correct them. **Figures 8**, **9**, and **10** illustrate examples for each case, while
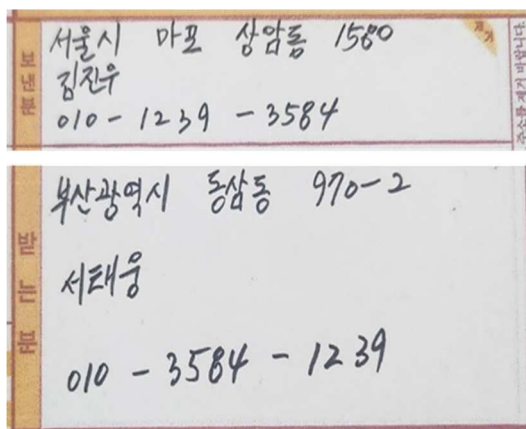


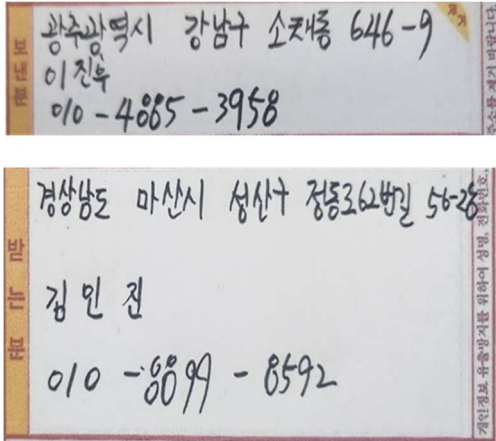**Figure 8:** Image of an Invoice with missing address

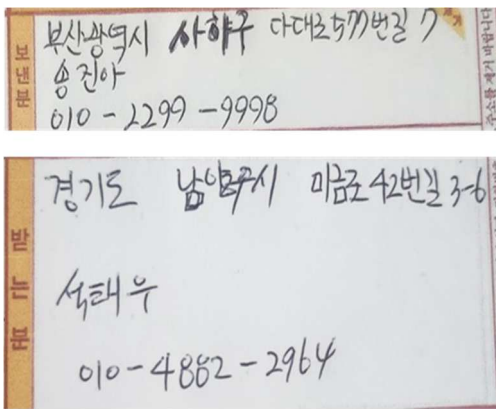**Figure 9**: Image of an invoice with address typos



**Figure 10**: Image of a Poorly Written Handwritten Invoice

**Table 3 :** Results of applying the model for Image of an Invoice with missing address

| Model | Sender | Receiver |
|---|---|---|
| GT | 서울 마포구 상암동 1580 | 부산광역시 영도구 동삼동 970-2 |
| Baseline[29] | 서울시 마포구 상암동 1580 | 부산광역시 동삼동 970-2 |
| Baseline + Ours | 서울특별시 마포구 상암동 1580 | 부산광역시 영도구 동삼동 970-2 |

**Table 4 :** Results of applying the model for Image of an invoice with address typos

| Model | Sender | Receiver |
|---|---|---|
| GT | 광주광역시 동구 소태동 646-9 | 경상남도 창원시 성산구 정동로62번길 56-28 |
| Baseline[29] | 광주광역시 강남구 소태동 646-9 | 경상남도 마산시 성산구 정동로62번길 56-28 |
| Baseline + Ours | 광주광역시 동구 소태동 646-9 | 경상남도 창원시 성산구 정동로62번길 56-28 |

**Table 5 :** Model application results for Image of a Poorly Written Handwritten Invoice

| Model | Sender | Receiver |
|---|---|---|
| GT | 부산광역시 사하구 다대로577번길 7 | 경기도 남양주시 미금로 42번길 3-6 |
| Baseline[29] | 부산광역시 사양구 다대로 577번길 7 | 경기도 남양시 미금로 42번길 3-6 |
| Baseline + Ours | 부산광역시 사하구 다대로577번길 7 | 경기도 남양주시 미금로 42번길 3-6 |

**Tables 3, 4, and 5** present the ground truth (GT) values, OCR predictions, and the results after applying the address correction model.

### 4.4 Discussion

In this paper, we measured the accuracy of invoice border detection and the address correction algorithm for handwritten invoice information extraction. For invoice border detection, the existing Yolact model was further trained with 10,000 additional box and invoice photos, and the accuracy of the Box Segmentation Module was evaluated using approximately 2,000 test datasets. The detection results showed high accuracy with 91.05% at mAP 50 and 87.02% at mAP 75. As a result, the Edge Detection Module was able to accurately separate sender and receiver images. To evaluate the address correction algorithm, we compared Model 1, which was trained using a certain percentage of generated data based on Donut, with Model 2, which included the address correction algorithm in Model 1. Compared to the original model, the number of correct addresses increased from 291 to 312, and the overall accuracy improved by approximately 5%.

According to the predictive analysis of the system, invoice detection was expected to be robust to both optical noise and physical noise, and the address correction algorithm was expected to produce an error of 10% or less for addresses only, by correcting misdetected words or typos in the results produced by the pre-trained model.

In the actual experiment, we found that the invoice detection was robust to various noise environments with high accuracy. However, in the case of the address correction algorithm, we found that the qualitative evaluation showed that some accuracy was improved by correcting some false detection words and typos only for addresses. However, the expected results were not achieved because it was difficult to distinguish between intentional typos and duplicate address names in different regions even with the address correction algorithm.

In conclusion, this study demonstrates the potential for improving the segmentation and VDU models used in unmanned parcel receivers for logistics system automation, and emphasizes that the accuracy of handwritten invoice recognition can be improved by additional methods such as LLM-based address correction algorithms. This provides important guidance for future model enhancement and commercialization strategies.

## 5. Conclusion

In this paper, we conducted a study on a handwritten invoice recognition system that improves invoice recognition performance through segmentation and enables address correction. The proposed system uses a segmentation model and an image processing algorithm to accurately crop the region containing information in the image, and modifies the address information outputted by the donut-based LLM appropriately to obtain higher accuracy results. In particular, we proposed a strategy for the advancement and commercialization of the entire system through the pre- and post-processing of the invoice region extraction method and the VDU model for handwriting recognition, which was inferred by the image processing algorithm alone. In the future, we will explore regularity and prediction methods for unstructured word or number arrays such as large-scale names and cell phone numbers to improve the model.

## Acknowledgement

## Author Contributions

Conceptualization, J. S. Lee and H. I. Seo; Methodology, J. S. Lee; Software, H. I. Seo; Validation, K. W. Han and H. I. Seo; Formal Analysis, K. W. Han; Data Curation, H. I. Seo; Writing Original Draft Preparation, J. S. Lee; Writing-Review & Editing, J. S. Lee. And H. I. Seo; Visualization, D. H. Seo; Supervision, Corresponding Author; Project Administration, D. H. Seo.

## References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," Proceedings of the IEEE international conference on computer vision, pp. 2961-2969, 2017.

[2] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," arXiv preprint arXiv:1609.02907, 2016.

[3] R. N. Khobragade, N. A. Koli, and V. T. Lanjewar, "Challenges in recognition of online and off-line compound handwritten characters: a review," Smart Trends in Computing and Communications: Proceedings of SmartCom 2019, pp. 375-383, 2020.

[4] S. Appalaraju, B. Jasani, B. U. Kota, Y. Xie, and R. Manmatha, "Docformer: End-to-end transformer for document understanding," In Proceedings of the IEEE/CVF international conference on computer vision, pp. 993-1003, 2021.

[5] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "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, 2020.

[6] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, and L. Zhou, "Layoutlmv2: Multi-modal pre-training for visually-rich document understanding," arXiv preprint, arXiv:2012.14740, 2020.

[7] Z. Yang, L. Li, J. Wang, K. Lin, E. Azarnasab, F. Ahmed, ..., and L. Wang, "Mm-react: Prompting chatgpt for multimodal reasoning and action," arXiv preprint, arXiv:2303.11381, 2023.

[8] S. Appalaraju, P. Tang, Q. Dong, N. Sankaran, Y. Zhou, and R. Manmatha, "DocFormerv2: Local Features for Document Understanding," arXiv preprint, arXiv:2306.01733, 2023.

[9] K. Lee, M. Joshi, I. R. Turc, H. Hu, F. Liu, J. M. Eisenschlos, and K. Toutanova, "Pix2struct: Screenshot parsing as pretraining for visual language understanding," In International Conference on Machine Learning, pp. 18893-18912, 2023.

[10] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9157-9166, 2019.

[11] V. Vuori, Adaptive Methods for On-Line Recognition of Isolated Handwritten Characters, Helsinki University of Technology, 2002.

[12] M. Ma, D. W. Park, S. K. Kim, and S. An, "Online recognition of handwritten Korean and English characters," Journal of Information Processing Systems, vol. 8, no. 4, pp. 653-668, 2012.

[13] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," In Proceedings of the IEEE International Conference on Computer Vision," pp. 569-576, 2013.

[14] Z. Zhang, C. Zhang, W. Shen, C. Yao, W. Liu, and X. Bai, "Multi-oriented text detection with fully convolutional networks," In Proceedings of the IEEE conference on computer vision and pattern recognition," pp. 4159-4167, 2016.

[15] A. Gupta, A. Vedaldi and A. Zisserman, "Synthetic data for text localisation in natural images," In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2315-2324, 2016.

[16] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "Layoutlmv3: Pre-training for document ai with unified text and image masking," In Proceedings of the 30th ACM International Conference on Multimedia, pp. 4083-4091, 2022.

[17] V. P. d'Andecy, E. Hartmann, and M. Rusiñol, "Field extraction by hybrid incremental and a-priori structural templates," 2018 13th IAPR International Workshop on Document Analysis Systems (DAS), pp. 251-256, 2018.

[18] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria, "Automatic semantics extraction in law documents," in Proceedings of the 10th international conference on Artificial intelligence and law, pp. 133-140, 2005.

[19] E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language, Springer, vol. 6036, 2010.

[20] J. Y. Lee and H. Y. Kim "A geocoding method implemented for hierarchical areal addressing system in Korea," Spatial Information Research, vol. 14, no. 4, pp. 403-419, 2006.

[21] J. Voerman, A. Joseph, M. Coustaty, V. P. d'Andecy, and J. M. Ogier, "Evaluation of neural network classification systems on document stream," In Document Analysis Systems: 14th IAPR International Workshop, DAS 2020, Wuhan, China, July 26–29, 2020, Springer International Publishing, Proceedings 14, pp. 262-276, 2020.

[22] A. Shapenko, V. Korovkin, and B. Leleux, "ABBYY: The digitization of language and text," Emerald Emerging Markets Case Studies, vol. 8, no. 2, pp. 1-26, Jun. 2018.

[23] F. Cesarini, E. Francesconi, M. Gori, S. Marinai, J. Q. Sheng, and G. Soda, "Conceptual modelling for invoice document processing," in Proceedings of Database and Expert Systems Applications. 8th International Conference, pp. 596-603, 1997.

[24] M. Köppen, D. Waldöstl, and B. Nickolay, "A system for the automated evaluation of invoices," in Document Analysis Systems(Series in Machine Perception and Artificial Intelligence), J. J. Hull and S. L. Taylor, Eds. Singapore: World Scientific, vol. 29, pp. 223-241, 1996.

[25] E. Medvet, A. Bartoli, and G. Davanzo, "A probabilistic approach to printed document understanding," International Journal on Document Analysis and Recognition, vol. 14, no. 4, pp. 335–347, Dec. 2011.

[26] I. Taleb, A. N. Navaz, and M. A. Serhani, "Leveraging large language models for enhancing literature-based discovery," Big Data and Cognitive Computing, vol. 8, no. 11, 146, 2024.

[27] V. Perot, K. Kang, F. Luisier, G. Su, X. Sun, R. S. Boppana, ..., and N. Hua, "Lmdx: Language model-based document information extraction and localization," arXiv preprint, arXiv:2309.10952, 2023.

[28] G. Kim, T. Hong, M. Hong, J. Park, J. Yim, W. Hwang, ... and S. Park, "Donut: Document understanding transformer without ocr," arXiv preprint, arXiv:2111.15664, vol. 7, no. 15, p. 2, 2021.