Journal of Advanced Marine Engineering and Technology, Vol. 46, No. 6, pp. 430~438, 2022
J. Advanced Marine Engineering and Technology (JAMET)
https://doi.org/10.5916/jamet.2022.46.6.430

ISSN 2234-7925 (Print)
ISSN 2765-4796 (Online)
Original Paper

A study on handwritten parcel delivery invoice understanding model

Yeong-Jae Shin[1] · Seong-Beom Jeong[2] · Hong-Il Seo[3] · Won-Yeol Kim[4] · Dong-Hoan Seo†

(Received December 21, 2022 ; Revised December 26, 2022 ; Accepted December 26, 2022)

Abstract: Optical character recognition (OCR) technology is a field of continuous research in which text images are stored or utilized as data. However, OCR technology alone has limitations in classifying and recognizing attribute values such as address, name, and phone number expressed in text in a semi-structured form, such as a parcel delivery invoice (PDI) written by hand. Therefore, in this study, we propose a handwritten parcel delivery invoice understanding (HPDIU) model for automated parcel delivery reception. The proposed HPDIU model consists of two steps: region detection of the parcel delivery invoice (RD-PDI) and information extraction from the PDI (IE-PDI). The RD-PDI, which is the first step, minimizes the resolution adjustment by detecting only the necessary area, including the sender and recipient information in the image. The second step, IE-PDI, consists of an end-to-end framework without OCR technology using a document understanding transformer and integrates the process of character detection, recognition, and understanding. In other words, the proposed model can solve the limitations of the OCR technology because it can integrate the process of classifying and recognizing according to attributes. To prove the validity of the proposed model, we used 500 handwritten PDI datasets to evaluate the accuracy of the character units according to the attributes of address, name, and phone number. As a result of the evaluation, a total average of 91.67% in units of letters proved the superiority of the proposed HPDIU model.

Keywords: Optical character recognition, Handwritten parcel delivery invoice understanding, End-to-End framework, Document understanding transformer

1. Introduction

The demand for non-face-to-face courier reception systems has significantly increased as the market size related to logistics transportation has recently exploded **[1][2]**. The unmanned courier reception system is being expanded and introduced because it can reduce the workload of employees and the cost of mail processing by shortening the waiting time for the convenience and accessibility of users. To fill out parcel delivery invoice (PDI) information, it is divided into two methods: writing with a digital device and writing by hand. The method of writing with a digital device can efficiently reduce the waiting time because the address can be entered through an unmanned reception machine. However, digitalization makes it difficult for the elderly or information-vulnerable to use it. Technical considerations are required to protect the underprivileged in a changing digital society. In addition, the handwriting method is problematic in that it is difficult to reduce the waiting time because the employee re-enters the address information written by the user into the reception system to receive the parcel. Therefore, technology that automatically recognizes PDI information, such as address, name, and phone number, is a core research technology that is very important in terms of economy and industry.

As PDI is composed of characters, it is divided into two technologies **[3]**– a technology for recognizing characters and one for

† 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 B. S., Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: yjshin0329@gmail.com, Tel: 051-410-4822

2 B. S., Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: sincere96@gmail.com, Tel: 051-410-4822

3 M. S., Department of Electrical & Electronical Engineering & Interdisciplinary Major of Maritime AI Convergence, Korea Maritime & Ocean University, E-mail: seoluck77@gmail.com, Tel: 051-410-4822

4 Ph. D., Artificial Intelligence Convergence Research Center for Regional Innovation, Korea Maritime & Ocean University, Email: kwy00@g.kmou.ac.kr, Tel: 051-410-4822

classifying the properties of recognized characters. OCR technology, which recognizes characters, is divided into print and cursive recognition **[4]**. According to the characteristics of standardized printed materials, the OCR technology for printed materials has been commercialized and generalized. However, OCR technology for handwriting is commercialized in limited situations because it is affected by individual handwriting, the condition of the fan, and even the material of the paper. In addition, English cursive recognition can be performed by classifying 52 letters of the alphabet by combining uppercase and lowercase letters. However, because Korean cursive recognition is composed of consonant and vowel combinations in syllable units, the total number of characters to be classified is 11,172 and the classification difficulty is very high. In addition to these problems, OCR technology alone cannot understand the meaning of text or classify attributes.

Recently, research on visual document understanding (VDU), a technology that can understand the properties of recognized characters, has been actively conducted along with the development of OCR technology. VDU can be utilized for various applications, including document classification **[5][6]**, information extraction **[7][8]**, and visual question answering **[9][10]**. There is a strong advantage in improving the efficiency of tasks by automating the processing of document images such as commercial invoices, receipts, and business cards that require manual work. In addition, the range of applicable industries is wide, so research based on deep learning is rapidly expanding. VDU additionally performs the process of receiving and understanding the text recognized through OCR. Because these technologies are generally OCR-dependent, they have some fatal drawbacks. First, the OCR techniques used for pre-processing are expensive because they require training costs and large datasets. The second is the negative impact of incorrect OCR on the VDU technology. Post-OCR correction module **[11]-[13]** technology is generally used to process this. However, this increases the overall system size and maintenance cost; therefore, it is not a practical solution for real application environments.

To solve this problem, this study proposes a transformer-based handwritten parcel delivery voice understanding (HPDIU) model. The proposed HPDIU model consists of two steps: detecting the parcel delivery invoice (PDI) area in the image and extracting the PDI information. In the first step, region detection of PDI (RD-PDI) minimizes resolution adjustment by detecting only the necessary area, including the sender and recipient information in the image. The second step, information extraction of PDI (IE-PDI), consists of an end-to-end framework that integrates character detection, recognition, and understanding processes without OCR technology using a document understanding transformer (Donut) **[14]**.

The proposed model has three strengths. First, the proposed model can recognize cursive and classify names, addresses, and phone numbers by attributes, thereby reducing the time to fill in addresses and personal information during unmanned parcel delivery. Second, the IE-PDI of the proposed model can reduce the dataset construction time because it does not require labeling for text-area detection. There is a limit to separately building each dataset to train the existing OCR and VDU models. Because the proposed HPDIU model integrates the two processes, a dataset for text-area detection is unnecessary. Third, the proposed model can reduce the amount of computation in the PDI information extraction step because it can remove unnecessary regions from the images through PDI region detection.

2. Related works

2.1 Optical Character Recognition

OCR is a traditional research topic in the field of image processing and refers to machine recognition of printed or handwritten characters. In other words, it is a technology for recognizing the content corresponding to a character in a document image input through a scanner to develop a digital document system. In the early stages, template matching and statistical and structural analysis methods were studied. There is a disadvantage in that the recognition time may be long because the rules for feature characters vary greatly depending on the font. Recently, studies using artificial neural network (ANN) models have mainly been conducted to recognize character patterns. Shi *et al.* **[15]** proposed a CRNN, which is a combination of a CNN and RNN. It has the advantage of being able to learn directly from words rather than letters. As a result, they use less storage space because they contain far fewer parameters than the standard DCNN models. In addition, Shi *et al.* **[16]** presented a spatial transformer network that can consider irregular shapes, such as perspective distortion and letter placement in words in natural images, unlike words in documents. Unlike using a separate text correction element in general, the recognition accuracy of irregular scene text was greatly improved by using attention-based spatial transformation network modules. Because the general OCR technology

reads the entire document and recognizes characters, unnecessary parts may be recognized. This sometimes causes inconvenience because it must be manually sorted to solve it. Therefore, it is important to extract only the necessary information by recognizing and understanding the text in a document.

2.2 Visual document understanding

VDU refers to understanding and extracting essential information from structured documents, such as receipts, invoices, and business cards. To perform VDU, most methods consist of three steps. First, the text is extracted using OCR technology, and then the OCRed text is serialized into a sequence of tokens. Finally, if possible, it was fed into the language model along with some visual features. Hwang *et al.* **[17]** proposed a spatial dependency parser to solve the problem of not being able to handle complex spatial relationships in structured documents easily. Improved accuracy in form comprehension tasks for real documents by formulating it as a problem of constructing a spatial dependency graph of documents. Hwang *et al.* **[18]** proposed an end-to-end framework based on serialization after OCR through weak supervision. As a result, there is a disadvantage that the performance is dependent on the performance of OCR. In addition, because OCR and VDU are separated, a large dataset must be built to improve performance through pre-training. Recently, Donut, an end-to-end framework from the OCR to VDU, was proposed **[14]**. Because the text processed by OCR is not required, the amount of computation is proven to be more effective than existing methods. However, because the processing process is performed without segmenting the input image, the recognition performance may deteriorate because the background image contains unnecessary information. In particular, the amount of calculation increases as the resolution has to be raised due to the background image.

3. Proposed HPDIU model

3.1 Overall architecture of proposed HPDIU model

The proposed HPDIU model aims to automate delivery receipts through manual invoice recognition. **Figure 1** shows the overall architecture of the proposed system, composed of region detection of PDI (RD-PDI) and information extraction of PDI (IE-PDI). The RD-PDI process converts an image containing only sender and recipient information from the courier box image scanned above. IE-PDI classifies the properties of the address, name, and phone number handwritten on PDI in the image

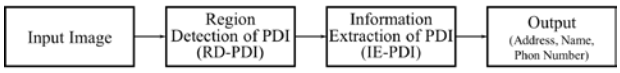


Figure 1: Overall architecture of HPDIU

detected through RD-PDI and converts the recognized result into text form. Because the proposed model does not require the labeling of text areas, it is easy to build a dataset. In addition, it is efficient because only the image region, including the property value, is inputted.

3.2 Region detection of PDI

Region detection in the PDI model detects the region where the sender and recipient are written in the image taken from the top of the delivery box in five steps, as shown in **Figure 2**. The first step is denoising to remove the noise caused by intrinsic or extrinsic factors of the image. The method uses nonlocal means denoising, which shows superior performance compared to Gaussian and median filters [19]. The denoised pixel value u(p) is obtained by dividing the sum of the products of the weight factor f(p, q) and the original image by the sum of the weight factors, as follows:

$$\mathrm{u(p)} = \frac{1}{C(p)} \sum_{q \in \Omega} v(q) f(p, q), \tag{1}$$

where p is the reference pixel, q is the surrounding pixel, Ω is the area of the image, and v(q) is the value of the original image at q. The weight factor is used to determine the value of a pixel with high similarity using the Euclidean distance between p and q, as shown

$$f(p, q) = e^{-\frac{|B(p)-B(q)|^2}{h^2}}, \tag{2}$$

B(p) is the average of the pixels around p and h is a filtering parameter. In the segmentation step, the image expressed in the RGB color space is converted to the HSV color space to intuitively check the color by expressing the image at an angle to reduce the effect of brightness. Then, because the color of the box is brown, the box and the background image are divided by converting the hue range from 0 to 20 into a black area. The segmented image detects the vertices of the invoice through the parcel delivery invoice edge detection step, which is composed of an edge detection algorithm and a contour detection algorithm. The detected vertex is a square rotated or distorted according to the shape attached to the package. Therefore, in the 2D image

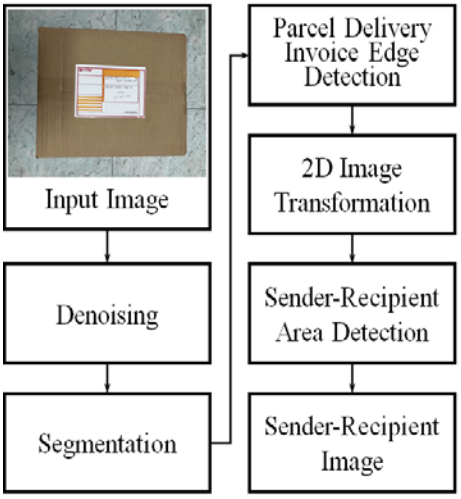


Figure 2: Structure of RD-PDI

transformation step, the distorted image is converted into a rectangular shape using homography to input a certain invoice shape into the OCR model. Finally, the sender and recipient areas were cropped using the coordinates of the invoice known in advance.

3.3 Information extraction of PDI

The donut model was used to perform PDI information extraction. Compared to other VDU models, donut has the advantage of exhibiting a high accuracy performance with a relatively small amount of computation. The structure of the proposed model consists of a visual encoder for extracting spatial features from the image output through the region detection of the PDI model and a textual decoder module to output a structured format by mapping the extracted features to lower word tokens, as shown in **Figure 3**. Therefore, it can be easily learned using an end-to-end method without an existing OCR. The visual encoder converts input image x into an embedding set z. We used a proven shifted windows (Swin) transformer **[20]** from Donut. The Swin converter divides the input image into patches of non-overlapping sizes through patch partitions and then makes them into features. Then, it is input to the Swin transformer through linear embedding. Each Swin transformer block is a structure in which windows multi-head self-attention (W-MSA) and shifted windows multi-head self-attention (SW-MSA) are continuously connected. W-MSA performs self-attention only between patches inside non-overlapping local windows. SW-MSA performs self-attention in a shifted window not only in a fixed location, but also in various areas. The final output of the Swin transformer is the input to the textual decoder.

The textual decoder converted the input visual encoder into

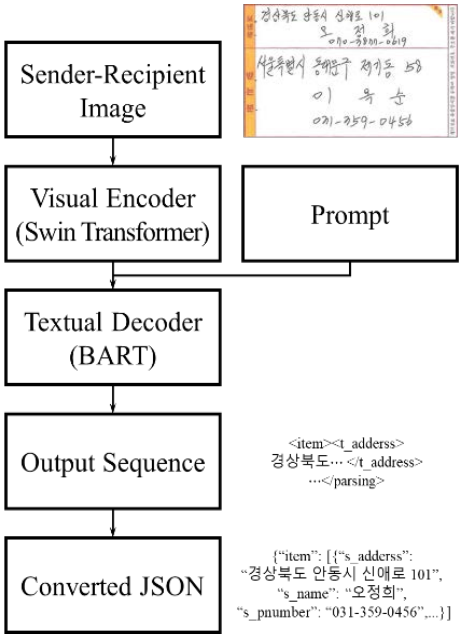


Figure 3: Structure of IE-PDI

a sequence of tokens. We used a bidirectional autoregressive transformer (BART), whose performance has been proven in Donut for textual decoding **[21]**. BART is a denoising autoencoder built on the seq2seq model, which has a bidirectional encoder, as in the BERT model **[22]**, and an auto-regressive decoder, as in the GPT model **[23]**. In addition, BART can apply any arbitrary transformation directly to the existing text, even by changing its length. Consequently, the sequence structure output by the textual decoder is in the form of a markup language that wraps elements with tags. The tag that expresses the properties of PDI consists of six tags (name and address, name, and phone number of the sender and receiver). Finally, the results output by the textual decoder are converted into JSON format so that they can be easily parsed and created in the software.

4. Experimental Environment and Evaluation

4.1 Dataset and training environment

The dataset used in the experiment consisted of 500 handwritten invoices and labels written by 20 middle-aged people in Yeongdo-gu and five students from the Korea Maritime and Ocean University for training and evaluating the HPDIU models. Because the properties of the courier invoice, such as address, name, and phone number, are private information, labels consisting of attribute values and words were created using the Faker library. The properties of the courier invoice were written in post office forms. An address label consists of 3–6 words, 3 syllables,

and a phone number of 9–11 syllables.

The environment settings of the server used for the neural network were as follows: Torch 1.11.0+cuda11.3 and torchvision 0.12.0 +cuda11.3 were used in the training framework, and an Nvidia RTX 8000 was used as the GPU. During training, the Adam optimizer with a learning rate of 0.00002 was used, and the validation data was set to 10% of the training data. The neural network was trained for 50 epochs in total, and the batch size was 8.

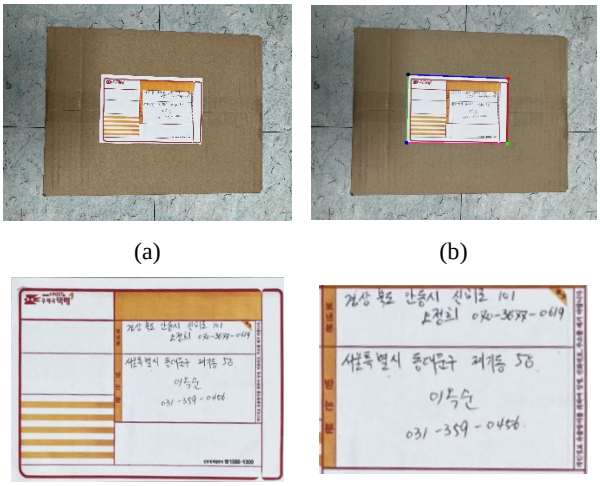
4.2 Evaluation method

Two experiments were conducted to evaluate the effectiveness of the proposed PDIUT. First, the detection process is performed from the original image to the image containing the information of the sender and recipient. The original video used here was an image that included the input courier invoice and background video, and the input size was 4032 × 3024.

Second, the accuracy of each class of PDIUT is quantitatively evaluated using the BLEU score, and the actual output result is compared with the ground truth to perform a qualitative evaluation. Because the number of datasets to be trained was small, fine-tuning was performed based on the donut-based model provided by Naver Clover, where pre-training was performed. The donut-based model is trained with a dataset of 11 M English document images provided by IT-CDIP **[24]** and a synthesized dataset of 0.5 M for each language using Wikipedia in Chinese, Japanese, Korean, and English. The size of the original image was 1,200 × 900, and when inputting it to the model, it was resized to 960 × 720 to reduce the amount of computation of the model. In the Swin-B transformer model used as a video encoder, the number of layers was 2, 2, 14, and 2, and the window size was 10.

5. Experimental Results

Figure 4 shows the ROI detection process for the postal parcels. **Figure 4** (a) shows an image entered into the postal address information recognition system and is similar to a kiosk environment that accepts parcels. **Figure 4** (b) shows the results of detecting the area of the parcel delivery invoice through denoising, segmentation, and edge detection, and it can be confirmed that the edge of the invoice is accurately detected. **Figure 4** (c) shows a parcel delivery invoice that outputs the detected area through a 2D transformation, and **Figure 4 (d)** shows an image obtained by extracting only the sender-recipient area using prior information



(a) (b) (c) (d)

Figure 4: Process of region detection of PDI

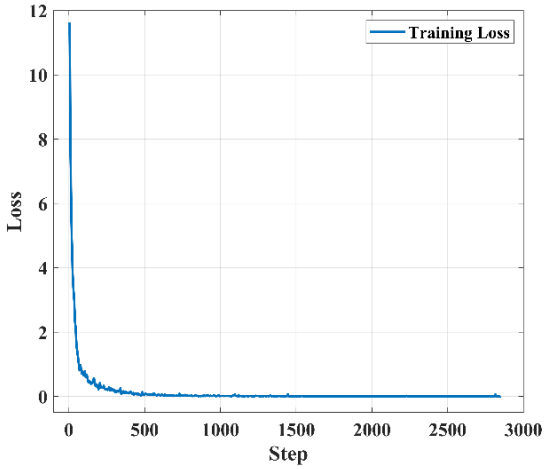


Figure 5: Training loss of HPDIU

on the invoice. In the experiment, it was confirmed that the preprocessing module detected the parcel delivery sender-recipient area and converted it into the minimum input data required by PDIUT.

Figure 5 shows the loss value according to the step to show the results of model training. As for the loss in the first step, it can be seen that the difference between the pretrained model and the prediction result is quite large. As the steps progressed, the loss gradually stabilized, and the difference was insignificant from 500 or more. This is because the words to be recognized are Korean cursive, and the structure of the document is different from that of the previously learned dataset. Despite the differences in these datasets, stable fine-tuning results were observed.

Table 1 shows the results of the character-by-letter evaluation according to the properties of the courier invoice used to evaluate the quantitative performance of the proposed model. Here, the

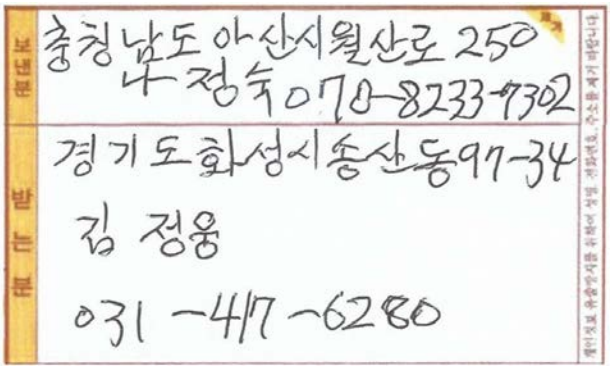
Table 1: Quantitative evaluation result of HPDIU

	Scale	Address	Name	Phone number
Total	-	1,947	300	1,210
Accuracy (%)	1/4	85.02	79.15	94.71
	1/3	89.25	83.97	97.19
	1/2	88.51	87.33	97.93
	1	88.03	91	97.93

properties of the sender and receiver were combined and evaluated. In addition, by evaluating the performance that can vary depending on the image size, the amount of computation and recognition accuracy can be evaluated by considering the image size, including the background image. We assume that, for most predictions, there are few out-of-order or duplicate characters. The experiment showed that the phone number matching accuracy was the highest at 97.93% and the address prediction result was the lowest at 88.03%. Errors are higher than for other attributes because addresses do not have a simple structure in the form of other phone numbers or names. In addition, it can be observed that the recognition accuracy for names and phone numbers decreased as the scale changed. By contrast, the accuracy of the address fell from 1/4 of the scale. We can see that the overall performance degrades significantly owing to a change in scale. The average accuracy of the original image was 91.76%, proving that a considerably high performance can be obtained by understanding the structure of the courier invoice using the proposed method. If the scale is converted to 1/4, we see a performance drop of approximately 3% to an average accuracy of 87.90%. Here, the result of the experiment that did not include the background image is depicted. Therefore, if the background and PDI are not distinguished, the difference in recognition accuracy according to scale may be further increased.

Figure 6 shows the results of the words generated by each model for a given image to compare whether the words generated by the proposed model are accurately provided. The words in the JSON format generated by the proposed model were compared with ground truth words. It can be seen that the attributes of the address, name, and phone number were well detected for each image. In the sender information in **Figure 6(a)**, even if the name and phone number are on one line, they are separately detected, and it can be confirmed that the attribute is correctly matched. In **Figure 6(b)**, even though the font is different from **Figure 6(a)**, the correct word was generated for each attribute.

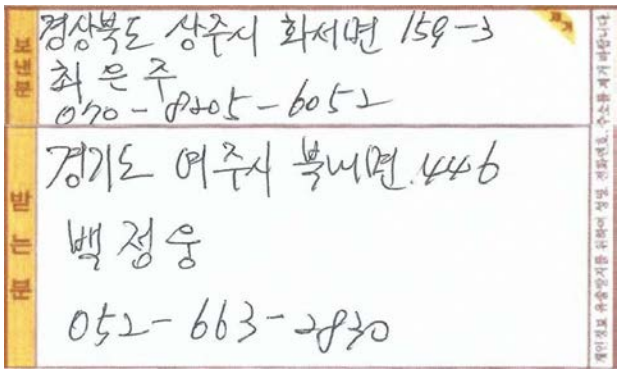
These results prove that correct words can be accurately generated for each attribute, despite the difference in fonts for each person because of cursive handwriting. However, because it is a cursive



(a)

Proposed Model: {'s_pnumber': '078-8233-7302', 's_name': '나정숙', 's_address': '충청남도 아산시 월산로 250', 'r_pnumber': '031-417-6280', 'r_name': '김정웅', 'r_address': '경기도 화성시 송산동 97-34'}

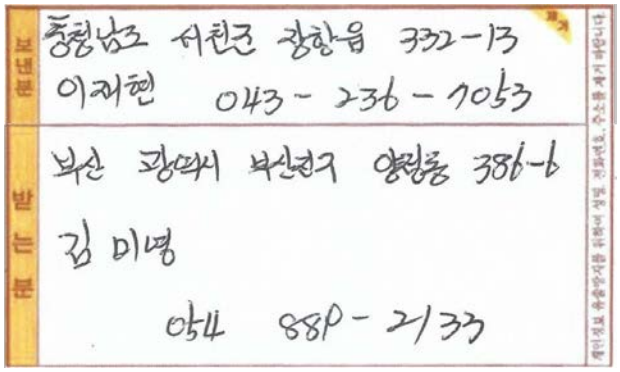
Ground truth: {'s_pnumber': '070-8233-7302', 's_name': '나정숙', 's_address': '충청남도 아산시 월산로 250', 'r_pnumber': '031-417-6280', 'r_name': '김정웅', 'r_address': '경기도 화성시 송산동 97-34'}



(b)

Proposed Model: {'s_pnumber': '070-8205-6052', 's_name': '최은주', 's_address': '경상북도 상주시 화서면 159-3', 'r_pnumber': '052-663-2830', 'r_name': '백정웅', 'r_address': '경기도 여주시 북내면 446'}

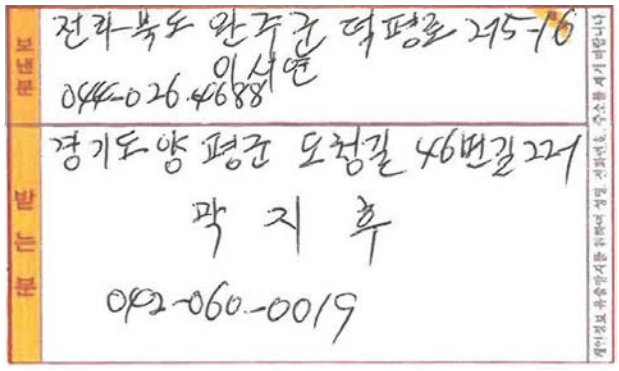
Ground-truth: {'s_pnumber': '070-8205-6052', 's_name': '최은주', 's_address': '경상북도 상주시 화서면 159-3', 'r_pnumber': '052-663-2830', 'r_name': '백정웅', 'r_address': '경기도 여주시 북내면 446'}



(c)

Proposed Model: {'s_pnumber': '043-236-7053', 's_name': '이재현', 's_address': '충청남도 서천군 장창군 332-13', 'r_pnumber': '054-889-2133', 'r_name': '김미영', 'r_address': '부산광역시 부산진구 양평동 386-6'}

Ground truth: {'s_pnumber': '043-236-7053', 's_name': '이재현', 's_address': '충청남도 서천군 장항읍 332-13', 'r_pnumber': '054-889-2133', 'r_name': '김미영', 'r_address': '부산광역시 부산진구 양정동 386-6'}



(d)

Proposed Model: {'s_pnumber': '044-026-4688', 's_name': '임식윤', 's_address': '전라북도 완주군 덕평로 275-16', 'r_pnumber': '092-060-0019', 'r_name': '박지후', 'r_address': '경기도 양평군 오청길 46 번길 22-1'}

Ground truth: {'s_pnumber': '044-026-4688', 's_name': '이서연', 's_address': '전라북도 완주군 덕평로 275-16', 'r_pnumber': '042-060-0019', 'r_name': '곽지후', 'r_address': '경기도 양평군 도청길 46 번길 22-1'}

Figure 6: Qualitative evaluation results of HPDIU

font, the difference in height between characters or lines is unclear, unlike printed fonts, resulting in errors. **Figure 6(c)** shows an image in which character recognition failed because the spacing between the characters was too narrow. Because of the nature of handwriting, errors may occur if each letter overlaps. In addition, **Figure 6(d)** shows an image in which character recognition fails because the height interval between lines is too narrow. As the name and phone number overlapped, it was recognized as a species that did not exist in the name.

A large amount of data is required for more accurate recognition. In other words, the number of words that can be expressed in Hangeul is 11,172, and the number of cases is inevitably large because the cursive style of Hangeul is diverse owing to individual differences. However, 450 data points were used to train the model, demonstrating that considerable accuracy could be obtained even with a fairly small dataset. In addition, because it is possible without a bounding box that indicates the position of a character, such as OCR, the time to build a dataset can be significantly reduced.

6. Conclusions

In this paper, we propose a preprocessing and artificial neural network model that can automatically extract address information from postal invoice images to reduce delivery time and user waiting time. Using Donut, an end-to-end framework that combines OCR and VDU, it is demonstrated that the process of detecting, recognizing, and understanding characters in postal invoices can be integrated. The proposed model has the advantage that the amount of computation is smaller than that of other methods that combine the OCR and VDU algorithms. In addition, unnecessary information was removed from the image through the preprocessing model so that the image, including the postal invoice and the background, could be considered.

In future, to utilize the proposed model for unmanned postal receivers, it should be designed as an embedded system. Therefore, we plan to conduct research on ways to simultaneously reduce weight and maintain performance at the same time.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00092, Development of core technology for delivery and parcel reception automation system using cutting-edge DNA and interaction

technology).

This research was supported by Korea Basic Science Institute (National research Facilities and Equipment Center) grant funded by Ministry of Education.(grant No. 2022R1A6C101B738)

Author Contributions

Conceptualization, D. H. Seo; Methodology, Y. J. Shin and W. Y. Kim; Software, S. B. Jeong and H. I. Seo; Formal Analysis, S. B. Jeong; Investigation, S. B. Jeong; Resources, S. B. Jeong; Data Curation H. I. Seo and W. Y. Kim; Writing-Original Draft Preparation, Y. J. Shin and W. Y. Kim; Writing-Review & Editing, Y. J. Shin; Visualization, Y. J. Shin; Supervision, D. H. Seo; Project Administration, D. H. Seo; Funding Acquisition, D. H. Seo.

References

[1] D. Wang, P. Hu, J. Du, P. Zhou, T. Deng, and M. Hu, "Routing and scheduling for hybrid truck-drone collaborative parcel delivery with independent and truck-carried drones," IEEE Internet of Things Journal, vol. 6, no. 6, pp. 10483-10495, 2019.

[2] E. Tallyn, J. Revans, E. Morgan, K. Fisken and D. Murray-Rust, "Enacting the last mile: Experiences of smart contracts in courier deliveries," Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-14, 2021.

[3] T. T. H. Nguyen, A. Jatowt, M. Coustaty, and A. Doucet, "Survey of post-OCR processing approaches," ACM Computing Surveys (CSUR), vol. 54, no. 6, pp. 1-37, 2022.

[4] N. Islam, Z. Islam, and N. Noor, "A survey on optical character recognition system," arXiv preprint arXiv:1710.05703, 2017.

[5] L. Kang, J. Kumar, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for document image classification," 2014 22nd International Conference on Pattern Recognition, pp. 3168-3172, 2014.

[6] M. Z. Afzal, S. Capobianco, M. I. Malik, S. Marinai, T. M. Breuel, A. Dengel, and M. Liwicki, "Deepdocclassifier: Document classification with deep convolutional neural network," 13th International Conference on Document Analysis and Recognition (ICDAR), pp. 1111–1115, 2015.

[7] W. Hwang, S. Kim, M. Seo, J. Yim, S. Park, S. Park, J. Lee, B. Lee, and H. Lee, "Post-OCR parsing: Building simple and robust parser via bio tagging," Workshop on Document Intelligence at NeurIPS, 2019.

[8] B. P. Majumder, N. Potti, S. Tata, J. B. Wendt, Q. Zhao, and M. Najork, "Representation learning for information extraction from form-like documents," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6495-6504, 2020.

[9] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A dataset for VQA on document images," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2199-2208, 2021.

[10] R. Tito, M. Mathew, C. V. Jawahar, E. Valveny, and D. Karatzas, "ICDAR 2021 competition on document visual question answering," International Conference on Document Analysis and Recognition, pp. 635-649, 2021.

[11] R. Schaefer and C. Neudecker, "A two-step approach for automatic OCR postcorrection," Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, pp. 52-57. 2020.

[12] S. Rijhwani, A. Anastasopoulos, and G. Neubig, "OCR post correction for endangered language texts," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5931-5942, 2020.

[13] Q. Duong, M. Hämäläinen, and S. Hengchen, "An unsupervised method for OCR post-correction and spelling normalisation for Finnish," Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa), pp. 240–248, 2020.

[14] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, and S. Park, "OCR-free document understanding transformer," European Conference on Computer Vision, pp. 498-517, 2022.

[15] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 11, pp. 2298-2304, 2017.

[16] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4168-4176, 2016.

[17] W. Hwang, J. Yim, S. Park, S. Yang, and M. Seo, "Spatial dependency parsing for semi-structured document

information extraction," Findings of the Association for Computational Linguistics, pp. 330–343, 2021.

[18] W. Hwang, H. Lee, J. Yim, G. Kim, and M. Seo, "Cost-effective end-to-end information extraction for semi-structured document images," Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3375-3383, 2021.

[19] A. Buades, B. Coll, and J. -M. Morel, "Non-local means denoising," Image Processing On Line, vol. 1, pp. 208-212, 2011.

[20] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9992-10002, 2021.

[21] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," pp. 7871-7880, 2020.

[22] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[23] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018.

[24] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard, "Building a test collection for complex document information processing," Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 665–666, 2006.