The System for Recognizing Useful Information of the Client's ID-Card Based on Machine Learning Technologies

Oleksii Bychkov [1], Liudmyla Zubyk [1], Dmytro Gololobov [2], Yaroslav Isaienkov [3], Ganna Grynkevych [4] and Anastasiia Ivanytska [1]

[1] *Taras Shevchenko National University of Kyiv, 60 Volodymyrska str., Kyiv, 01601, Ukraine*
[2] *National Aviation University, 1 ave. Huzar Lubomyr, Kyiv, 03058, Ukraine*
[3] *Vinnytsia National Technical University, 95 Khmelnytske highway, Vinnytsia, 21021, Ukraine*
[4] *State University of information and communication technologies, 7 Solomyanska str., Kyiv, 03110, Ukraine*

Abstract

Existing approaches to recognition of text information from user documents in client-server systems were analyzed in order to solve the problem of user identification. The comparative analysis of the Optical Character Recognition library using Tesseract Engine and Paddle Optical Character Recognition was carried out. The feasibility and effectiveness of using Paddle Optical Character Recognition for analyzing documents with Chinese characters was substantiated. The text processing model was proposed to highlight valuable information and form the dictionary, which will be transmitted to the system server. Its effectiveness was verified on test data and the adequacy of the model was assessed based on Character Error Rate and Word Error Rate (CER and WER). In contrast to existing approaches, the model showed the increase in text recognition accuracy by 1.8-11.3%, depending on the quality of the source image. The result is implemented in client-server applications for product solutions of Zhejiang Jimi IoT Technology Co., Ltd.

Keywords [1]

Optical character recognition, machine learning, client-server systems.

1. Introduction

Identification of relevant information from the documents of the customer system and further identification of the system is the important part of the online applications of accounts and services.

Proposed extracting useful data from customer IDs card using Optical Character Recognition (OCR). After analyzing the literature [1-3], it was revealed that the Tesseract Engine library was used to recognize the texts of past documents (papers, articles) from popular libraries, tax card.

Tesseract, the open-source package implemented in Python, was used to process digitized images in the issue [4]. The OCR output was processed using Python modules applying localization and text detection followed by classification, but without the use of machine learning/deep learning/natural language processing techniques, which are quite complex and time and data intensive. However, the average accuracy of the obtained result was quite low (34.60%).

ID verification is undoubtedly one of the most difficult steps in the Know Your Customer (KYC) process, requiring the lot of effort, time and money to implement as usually. A revolutionary solution to ID verification problems involving machine learning and deep neural network techniques was proposed in the study [2].

The study [1] looked at the OCR model (Tesseract) built on CNN that extracted data from the tax card image. Good results were obtained, providing an accuracy level of 80%.

Dynamical System Modeling and Stability Investigation (DSMSI-2023), December 19-21, 2023, Kyiv, Ukraine
EMAIL: oleksiibychkov@knu.ua (O. Bychkov); zubyk.liudmyla@knu.ua (L. Zubyk); gololobov.dma@meta.ua (D. Gololobov); yisaienkov@gmail.com (Y. Isaienkov); Ggrynkevych@ukr.net (G. Grynkevych); a.titova.wk@gmail.com (A. Ivanytska)
ORCID: 0000-0002-9378-9535 (O. Bychkov); 0000-0002-2087-5379 (L. Zubyk); 0000-0003-4732-0071 (D. Gololobov); 0009-0005-5629-0021 (Y. Isaienkov); 0000-0003-1922-5165 (G. Grynkevych); 0000-0002-4803-2090 (A. Ivanytska)

OCR is the technology to identify characters with the highest possible accuracy through the use of appropriate pre-processing, processing and post-processing refinements. The significant contribution to the final result is provided by the avoidance of noise in the original image, which is emphasized in the study [5].

Automatic information extraction from scanned documents significantly increases efficiency, accuracy and speed in all those business processes where data collection from documents plays an important role. Such documents are often digitized as images and then converted to text format. A text recognition method based on transfer learning and scene text recognition (STR) networks is presented in [6-7]. The formalization model of recommendations that based on technology of machine learning was proposed. After analyzing of different methods, the product-based collaborative filtering (CF) was chosen to solve the research problem [8-9].

For experimental research the algorithm was used that underlies in information technology of person identification in video stream. The algorithm performance provided various identification accuracy rate results with the difference, which amount to 20% on average [10]. The block diagram and techniques for optimal implementation of the improved gradient method in device control for unmanned aerial vehicle are available in the paper [11].

In this paper [12], was proposed the new original architecture of a model based on an artificial convolutional neural network and semantic segmentation approach for the recognition and detection of identity documents in images. Authors [13] proposed a new pre-processing approach consists of DeblurGAN (Generative Adversarial Network for deburring image), shadow removal, and binarization to pre-process the image for Tesseract-OCR, achieved an average Character Error Rate of 18.82% which is better compared to without pre-processing which is 38.13%.

However, existing approaches to text recognition often require image pre-processing, which requires the development of additional models for processing the resulting text or input image. The main problem in text recognition is the quality of the input image, its brightness, position relative to the camera and size. Thus, today an urgent scientific task is to find approaches to image recognition of user documents, which provides the convenient format for the resulting data. To solve this problem, it is necessary to analyze different approaches to recognize documents with Chinese characters and develop a new model for forming the dictionary of Chinese user's information, which can be verified on test images.

2. Purpose of the study

The purpose of this research is to develop the model for forming the dictionary based on text recognition technologies from document images and increase the accuracy of information transfer in the client-service system. To achieve this goal, it is necessary to solve the following tasks:

- analysis of text recognition technologies on documents;
- select the optimal module from two approaches for recognizing text images
- determine the stages of developing the model of the dictionary formation which based of text recognition technologies, which was selected;
- evaluate the adequacy of the proposed model on test data.

2.1 Text recognition tools based on machine learning technologies

Initially, Tesseract technology was used during the research to recognize documents with Chinese characters. It can be used directly or through the API to extract written text from images with different languages. Tesseract may be used in conjunction with the existing layout analysis to recognize text inside a big document, or with an external text detector to recognize text from an image of a single text line. Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others formats of file [7, 14-16].

The first approach, Tesseract technology was used in the next steps, which presented in fig. 1. Test images were uploaded to the recognition system and recognition results were obtained. For good quality images taken directly, the system produced recognition results. To track the quality of the

recommendations of viewed methods, it is proposed to choose metrics from the following: root mean square error (RMSE), mean absolute error (MAE), normalized value of the mean absolute error (NMAE) [17-21]. But for evaluating the quality of recognition according to follow the parameters CER and WER, calculated according to expressions 1-2, the following results were obtained:

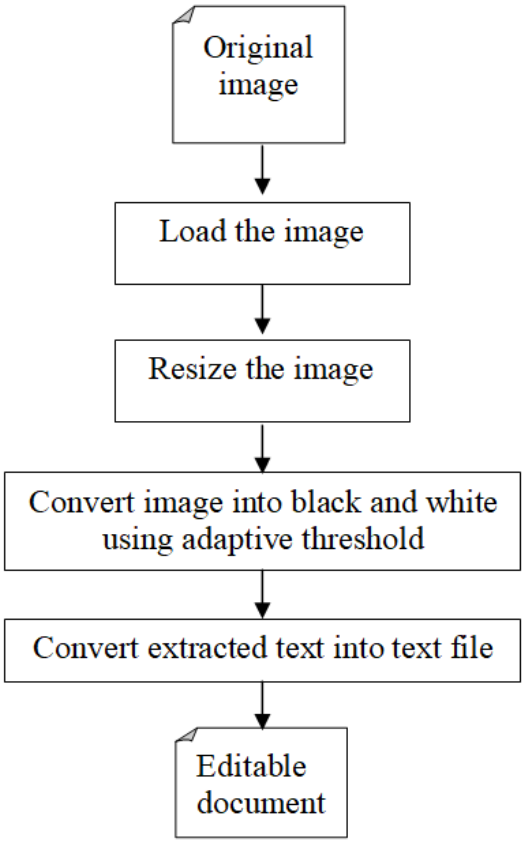

Figure 1: Sequence of data processing recognition system

$$CER = \frac{S + D + I}{N} \quad (1)$$

where S – number of substitutions;
D – number of deletions;
I – number of insertions;
N – number of characters in reference text.
$N = S + D + C$ (where C – number of correct characters).

$$WER = \frac{S_w + D_w + I_w}{N_w} \quad (2)$$

The formula for WER is the same as that of CER, but WER operates at the word level instead. It represents the number of word substitutions, deletions, or insertions needed to transform one sentence into another [22].

The CER becomes in interval [25%, 33.33%] and the WER value of 50% is clearly understood since 2 out of 4 words in the sentence were wrongly transcribed. The obtained results do not make it possible to build a text processing model for dictionary formation. Therefore, a decision was made to choose other technologies for conducting research, namely Paddle OCR.

Paddle OCR is the open-source OCR toolkit developed by PaddlePaddle, an advanced AI model based on the powerful GPT-3.5 architecture, developed by Open AI. As the cutting-edge OCR technology, Paddle OCR excels in converting images containing textual content into editable, searchable, and machine-readable text. It efficiently analyzes images and recognizes characters,

numbers, and symbols, enabling seamless image-to-text conversion. Paddle OCR consists of an ultra-lightweight and general OCR model, integrating OCR algorithms like:

- Text detection models: EAST, DB, SAST
- Text recognition models: CRNN, Rosetta, STAR-Net, RARE, SRN [8, 23-25].

The next machine learning technologies are used for the proposed text recognition system, such as Convolutional Recurrent Neural Network (CRNN), sequence recognition network (SRN). The network architecture of CRNN consists of three components, including the convolutional layers, the recurrent layers, and a transcription layer, from bottom to top. At the bottom of CRNN, the convolutional layers automatically extract a feature sequence from each input image. It can be applied to problems – Chinese character recognition, which involve sequence prediction in images. In this case, the following steps were used to find useful information from user documents:

- image loading;
- download of all libraries, packages and language configuration (chinese_cht.ttf);
- setting the recognition threshold to 0.1;
- text recognition from user documents.

Next, the Paddle OCR also was tested on 10 test images, and the system produced recognition results. For 6 images, the recognition result was obtained with the recognition accuracy in the range from 0.835 to 0.999.

For images of low quality, the system produces result that was difficult to process and reveal valuable information for further dictionary formation. After evaluating the quality of recognition according to the CER and WER parameters, calculated according to expressions 1-2, the next values were obtained, which are presented in Table 1.

Table 1

The quality of recognition according to the CER and WER parameters with address date of Chinese user's information

File name	OCR address	Real address	CER, %	WER, %
8-1.jpg	'广州市南沙区横沥镇大安 北街77号	'广州市南沙区横沥镇大安 北街77号	98.0	2.0
6-1.jpg	广州市番禺区洛浦街东乡 四街四巷七横巷2号	广州市番禺区洛浦街东乡 四街四巷七横巷2号	98.7	1.8
7-1.jpg	广州市番禺区洛溪新**东**吉 祥道一幢之二601房	广州市番禺区洛溪新城吉 祥道一幢之二601房	96.7	3.3
1.jpeg	安徽省涡阳县西阳镇**大**庙 行政村王大庄自然村049 号	安徽省涡阳县西阳镇文庙 行政村王大庄自然村049 号	95.8	4.3
4.jpg	河北省**庄**市郭村乡北**自**平庄村049号	河北省河间市郭村乡北太 平庄村049号	89.5	11.3

As can be seen from the table, the result of text recognition for images of good and normal quality (look at fig. 2) is sufficient for forming the text processing model and forming the dictionary for the client-server system. Characters that were incorrectly recognized are highlighted in bold. Also, the structure of the source text corresponds to the structure of the document, which greatly complicates the process of developing the dictionary formation model.

After a comparative analysis of the two approaches, Paddle OCR was selected for further research based on the text recognition quality scores for 10 images of varying quality.

Figure 2: The result of text recognition by the Paddle OCR system

2.2 Development of the text processing model for dictionary formation

The object of the study is the image of the client's ID card, which is presented in figure 2, after text recognition by the Paddle OCR system, the processing model was developed, which will allow the formation of the dictionary of useful information, which looks like this:

"name":"张*舒",
"sex":"女",
"nation":"汉",
"birth":"197***512",
"address":"广州市番**洛溪新城吉祥道十幢之二601房",
"idcard":"441*******0427

After text recognition, the text information processing model is proposed and includes the next stages of forming the dictionary of useful information:

1. Combining the recognition result into one string
2. Function to find part of text
3. Find id_number after the characters '公民身份号码' in the client's passport.
4. Using of regular expression for search id_number
5. Search Data of birthday after symbols '出生','住址'
6. Find address after characters '住址', '公民身份号码'
7. Find id_number after '住址'
8. Formation of the result table styling.

The developed text processing model made it possible to obtain dictionaries containing the user's personal data, so they will not be included in the research results. This made it possible to proceed to the development of the client-service system for user interaction and a user application.

3. Structure of the project

The project contains various components, to use the HTTP protocol for interaction. To initiate the request to identify an ID card, the following is the sample Python request code. The responder sends the generated dictionary to the server if the text recognition module, the post OCR model for processing useful information, or the error in recognition are successful. Then the user can upload the better quality image again and try identification again. To form the responder sends for the client-server system, the next steps are proposed:

1. Read the input image.
2. Run the text recognition algorithm.
3. Use text processing model for dictionary formation.

5. Send valuable information to the server.
6. If the result of forming the dictionary is error, then send the result – error to the document.

The result was implemented in client-server applications for product solutions of Zhejiang Jimi IoT Technology Co., Ltd.

4. Conclusions

As the result of the research, the model of forming dictionary basing on text recognition technologies, that uses machine learning technologies, was developed and implemented. The following stages of research were performed.

Firstly, existing approaches to recognition of text information from user documents in client-server systems were analyzed in order to solve the problem of user identification. Two approaches were selected for text recognition algorithm.

Secondly, the comparative analysis of the Optical Character Recognition library using Tesseract Engine and Paddle Optical Character Recognition was held. The feasibility and effectiveness of using Paddle OCR for analyzing documents with Chinese characters was substantiated.

Thirdly, the model of the dictionary formation which based of text recognition technologies was developed and tested.

Finally, the adequacy of the model of the dictionary formation was evaluated. Its effectiveness was verified on test data and the adequacy of the model was assessed based CER and WER. In contrast to existing approaches, the model showed the increase in text recognition accuracy by 1.8-11.3%, depending on the quality of the source image. The result is implemented in client-server applications for product solutions of Zhejiang Jimi IoT Technology Co., Ltd.

5. References

[1] Qamar Uddin, Features Extraction of Tax Card by Using OCR Based Deep Learning Techniques, Master's thesis, University of Eastern Finland, Faculty of Science and Forestry, Joensuu School of Computing. Computer Science, 2021.

[2] Tuan Bui, Applications of Machine Learning in EKYC'S Identity Document recognition. Bachelor's thesis, South-Eastern Finland University of Applied Sciences, 2021.

[3] T. Nguyen, A. Jatowt, M. Coustaty, N. Nguyen, and A. Doucet, Deep statistical analysis of OCR errors for effective post-OCR processing, ACM/IEEE Joint Conference on Digital Libraries (JCDL), Jun. 2019, 29–38. doi: 10.1109/JCDL.2019.00015.

[4] Ch. Padole, U. Sh. Verma, Pr. Gujral, Mr. Kumar etc., Information Extraction from Visiting Cards Using OCR and Post-Processing in Python, International Journal of Scientific and Technical Research in Engineering (IJSTRE), Vol.7, ISSUE 05(2022), 1-7.

[5] C. Madan Kumar, M. Brindha, Text Extraction from Business Cards and Classification of Extracted Text Into Predefined Classes, Proceedings of International Conference on Computational Intelligence & IoT (ICCIIoT), 2018, 595-602.

[6] D. Arulanantham, S. Snekha, K. Logeshwaran, R. Nishanth, K. Lavanya, Utilizing OCR to Retrieve Text from Identity Documents, International Journal for Research in Applied Science & Engineering Technology (IJRASET), Vol. 11, Issue IV, Apr 2023, 349-363. doi: 10.22214/ijraset.2023.50090.

[7] James A., Seth A., Mukhopadhyay S.Ch. Iot Systems Design : Projects Based Approach, Springer, 2022, 227-279. doi : 10.1007/978-3-030-85863-6_12

[8] What is Paddle OCR? URL: https://www.plugger.ai/blog/what-is-paddle-ocr

[9] Ivanytska A., Zubyk L., Zubyk Y., Ivanov D. The advertising predictivon model based on machine learning technologies // CEUR Workshop Proceedings of the 8th International Conference "Information Technology and Interactions" (IT&I-2021), 1-2 Desember, 2021.

[10] V. Martsenyuk, O. Bychkov, K. Merkulova and Y. Zhabska, Exploring Image Unified Space for Improving Information Technology for Person Identification // IEEE Access, vol. 11, pp. 76347-76358, 2023, doi: 10.1109/ACCESS.2023.3297488.

[11] Dakhno, N., Barabash, O., Shevchenko, H., Leshchenko, O., & Dudnik, A. Integro-differential models with a K-symmetric operator for controlling unmanned aerial vehicles using a improved gradient method. // 2021 IEEE 6th International Conference on Actual Problems of Unmanned Aerial Vehicles Development (APUAVD) October, 2021. pp. 61-65.
[12] M. Kozlenko, V. Sendetskyi, O. Simkiv, N. Savchenko and A. Bosyi, Identity Documents Recognition and Detection using Semantic Segmentation with Convolutional Neural Network, Cybersecurity Providing in Information and Telecomunicationg Systems, 2021, Kyiv, Ukraine, 234-242 doi :10.5281/zenodo.5758182
[13] E. Zhang, V. A. Putra, G. P. Kusuma, Improving Optical Character Recognition Accuracy for Indonesia Identification Card Using Generative Adversarial Network, Journal of Theoretical and Applied Information Technology, 2022, Vol. No 8, 2424-2437.
[14] M. R. M. Ribeiro, D. Julio, V. Abelha, A. Abelha, and J. Machado, A comparative study of optical character recognition in health information system, International Conference in Engineering Applications (ICEA), Jul. 2019, 1–5. doi: 10.1109/CEAP.2019. 8883448.
[15] Ivanytska A., Zubyk L., Ivanov D., Domracheva K. Study of methods of complex data analysis based on machine learning technologies IEEE Internation Conference on Advanced Trends in Information Theory, ATIT, 2019D.
[16] Phan Van Hoai, H.-Th. Duong, V. Th. Hoang, Text recognition for Vietnamese identity card based on deep features network, International Journal on Document Analysis and Recognition (IJDAR), 24(1–2), 2021. doi : 10.1007/s10032-021-00363-7
[17] Ivanytska, A., Zubyk, L., Dudnik, A., Kurchenko, O., Berestov, D. Evaluation of the Effectiveness of Information Technology Methods for Processing Diagnostic Information Based on Complex Data // 2022 IEEE 3rd International Conference on System Analysis and Intelligent Computing, SAIC 2022 - Proceedings, 2022.
[18] Kutyrev A., Kiktev N., Kalivoshko O., Rakhmedov R. Recognition and Classification Apple Fruits Based on a Convolutional Neural Network Model. (2022) CEUR Workshop Proceedings, 3347, pp. 90 – 101. https://ceur-ws.org/Vol-3347/Paper_8.pdf
[19] N. S. Yusman, M. M. Ibrahim, Extracting information from identity card into electronic form using image processing technique, Proceedings of Innovation and Technology Competition (INOTEK), Vol.1(2021), 2021, Melaka, Malaysia, 135-136
[20] Kr. Olejniczak, M. Sulc, Text Detection Forgot About Document OCR, 26th Computer Vision Winter Workshop, Robert Sablatnig and Florian Kleber (eds.), Krems. Lower Austria, Feb.15-17, 2023, 1-7.
[21] D. Yan, S. Shen and D. Wang, Functional Structure Recognition of Scientific Documents in Information Science, Join Workshop of the 4th Extraction and Evaluation of Knowledge Entities from Scientific Documents and the 3rd AI + Informatics (EEKE-AII2023), June 26, 2023, Santa Fe, New Mexico, USA and Online, 10-12.
[22] K. Leung, Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER). URL: https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510
[23] F. Borisyuk, A. Gordo, V. Sivakumar, Rosetta: Large scale system for text detection and recognition in images, KDD'2018, London, United Kingdom. doi.org/10.475/123_4
[24] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, Errui Ding, Towards Accurate Scene Text Recognition with Semantic Reasoning Networks, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13-19 June, 2020, DOI:10.1109/cvpr42600.2020.01213
[25] Xinyan Zu, Haiyang Yu, Bin Li, Xianguang Xue, Towards Accurate Video Text Spotting with Text-wise Semantic Reasoning, Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23), 2023, 1858-1866.