Human Related Information Extraction from Chinese Archive Images

Xin Jin[1], Hangbing Yin[1], Xiaoyu Chen[2], Huimin Bi[1], Chaoen Xiao[1(✉)], and Yijian Liu[1]

[1] Beijing Electronic Science and Technology Institute, No. 7, Fufeng Street, Fengtai District, Beijing 100070, China
xcecd@qq.com

[2] Information Center of China North Industries, Group Corporation Limited, Beijing 100089, China

Abstract. With the rise of the information age, digitalization and paperless processes have become the norm in managing archive images. However, research on extracting and managing image content from archives is still in its early stages, and is primarily focused on recognizing fixed-format archive images. As a result, there is a lack of technology for extracting key personal information applicable to all types of archive images. To address this, we have identified two main tasks: extracting identity photos and key personal information. To ensure confidentiality of real data, we created a dataset that simulates certificate photo files. We then used a YOLOv5-based object detection network to train a model that can detect document photos in archive images. We also used a combination of PP-OCR text recognition and object detection to extract key information from document images.

Keywords: File image processing · Certificate photo extraction · PP-OCR · YOLOv5 · Key information extraction

1 Introduction

In recent years, with the continuous development of digital archiving, many archival institutions have entered the "post-digitization" stage. A large number of archives are saved in image or PDF formats, but their content has not been fully explored and developed. When document images have intelligent analysis and review functions, key content such as ID photos, names, gender, birthdates, and important dates can be intelligently extracted from digital files through image analysis, thereby reducing the workload of inspection personnel. Regarding the face photos, traditional face detection methods are difficult to extract ID photos: there are various forms of ID photos in archival materials, many photos are severely worn, or become black and white blocks after multiple copies. Face recognition cannot identify the range of ID photos containing information such as hair, shoulders, and background, which does not meet the needs of extracting and cropping ID photos. As for the OCR model extraction method, traditional OCR methods are difficult to extract key information from table images: under the influence of table

H. Lu and J. Cai (Eds.): ISAIR 2023, CCIS 1998, pp. 139–146, 2024.
https://doi.org/10.1007/978-981-99-9109-9_14

horizontal and vertical lines, it is easy to recognize the lines as numbers or text, and the recognition efficiency and accuracy are poor, especially for archives with handwritten fonts, the recognition effect is extremely poor. OCR alone cannot realize the need to centrally extract handwritten information images for handwriting comparison.

To solve the above problems, our main research content is as follows: (1) Construction of archival image dataset: based on real archival datasets, relying on image processing and other technologies, a dataset containing various personnel ID photos and a table image dataset containing personnel key information are constructed to meet the needs of subsequent training and optimizing models. (2) Proposed a method for detecting archival certificate photos: automatically extract portraits from archival materials, comprehensively consider the detection and segmentation of facial parts and other parts in certificate photos, recognize and extract certificate photos in the input image, and batch output/save/display the source images. (3) Proposed a key information extraction method: intelligently analyze form files and propose a method that combines table recognition, Chinese OCR, and object detection to extract key personnel information from document images, and output the structured content to CSV for summary storage. This article consists of four chapters: Section 1 is the introduction, which introduces the main problem that the experiment is trying to solve. Section 2 details the design and experiment of extracting archival ID photos. Section 3 details the specific implementation process of extracting key personnel information. Section 4 summarizes the work.

2 Photo Extraction

2.1 Scheme Design

This article proposes a document ID photo extraction technology based on object detection. The document images containing ID photos are labeled using the LabelImg annotation tool to generate YOLO-format annotation files, which constitute the ID photo detection dataset. Then, the YOLOv5 [1, 9] network is used to train the object detection model. In practical applications, the document images to be processed are placed in a specified folder, and the inference detection module is run to display all the ID photos in the document. Finally, based on the detection results, OpenCV is used to crop the ID photos from the document and output them to a specified folder for centralized storage. The design of the document ID photo extraction module is shown in the Fig. 1 below.

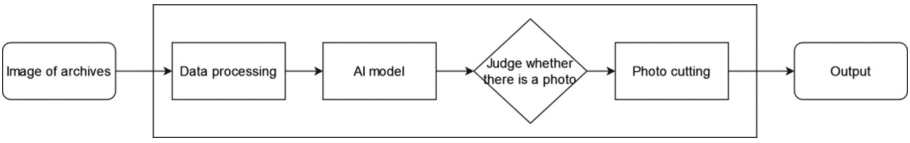

Fig. 1. Extraction technology of certificate photo.

2.2 Construction of Photo Dataset

To solve the problem of scarce real document data and diverse ID photo types, we propose a semi-automatic document image generation method with ID photos based on the real document image effect, which is used to construct a self-made ID photo document dataset. The document images containing ID photos can be divided into three categories: yellowed old documents, color documents, and black and white documents that have been copied. Based on these three categories, ID photos can be divided into five types: low-quality old photos, color photos, yellowed photos, black and white photos, and copied photos. In this section, we first use a web crawler to obtain corresponding color ID photos, and then use OpenCV to process the color space of the ID photos to obtain the effect of yellowing and black-and-white binarization. We also apply noise addition and smoothing to simulate the effect of multiple copies. The specific effects are shown in the Fig. 2 below.

Fig. 2. All kinds of homemade images.

2.3 Photo Extraction Model

Data Preparation and Model Training. In this paper, we utilized the open-source image annotation tool LabelImg to manually label seal images on the archive and obtain pixel information when synthesizing the homemade ID photo and real archive background. Then, when generating the composite archive, we directly generated the txt labels in YOLO [2] format according to the relative position of the ID photo in the archive and saved them. Automated label generation can greatly improve the efficiency of data annotation for creating a dataset, and manual review can be done later with the help of the LabelImg interface. We then chose YOLOv5s to train the model, trained it for 300 epochs, and obtained the ID photos [3, 8].

Image Cropping. During network training, YOLOv5 sets the initial anchor box size and outputs the predicted box based on it, and then updates the parameters of the network structure by propagating the error direction between the predicted box and the real box. During inference, the location of the ID photo is indicated and the type of the ID photo is displayed above the anchor box. Finally, the ID photo is clipped along the edge of the anchor box and the seal image is classified and stored in different folders.

2.4 Experimental Process and Analysis of Results.

The Verification about Validity of Self-made Data. We trained the model using a dataset of 698 real documents containing ID photos. The box_loss and obj_loss of the training set showed a decreasing trend overall, but their performance on the validation set was not stable and fluctuated greatly. Although the Precision and Recall of the model were both at a high level, the mAP@0.5:0.95 value was low and varied greatly, which affected the overall performance of the model. Therefore, the real document model was able to extract ID photos, but its performance needed improvement. On the other hand, the model trained using the self-made document dataset achieved good results, with GIoU and Objectness both at a low level. The Precision and Recall were close to 1, and the mAP@0.5 reached 0.99, while the MAP@0.5:0.95 reached 0.98. Therefore, the ID photo model had a high comprehensive recognition accuracy and was suitable for the document ID photo extraction module. After balancing the number of samples, the model's performance was greatly improved.

Validity Verification. We tested the real document model and self-made document model on real document and self-made document datasets, respectively. The self-made document model performed well on the self-made dataset, recognizing all ID photos with a recognition accuracy of 1. However, due to the richer information in real documents, although the self-made ID photo file model was able to recognize all ID photos during testing on a real dataset, there were still issues with over-detection and false detection. Therefore, we added some elements of misidentified images as negative samples in the self-made dataset to optimize it. By testing, we found that applying object detection to extract ID photos from documents was feasible. The next step would be to expand the dataset, optimize its composition, and improve the model's detection performance.

The Verification about Validity of Fusion Photo. The overall performance of the model fused with the archive data set is relatively stable, and the loss functions of the training set and the test set show an overall downward trend. Both Precision (accuracy rate) and Recall (recall rate) have reached close to 1. And mAP@0.5 reaches 0.99, and MAP@0.5:0.95 reaches 0.98. In summary, the ID card photo model has a high comprehensive recognition accuracy and is suitable for the file ID photo extraction module. The fusion photo model was tested on authentic data and homemade data, producing the corresponding metrics as presented in Table 1.

Table 1. Testing Results of Fused Identification Photo Model.

	Images	Labels	P	R	mAP@.5	mAP@.5:.95
Authentic Data	100	101	0.990	0.986	0.992	0.809
Homemade data	100	100	0.997	0.972	0.991	0.964

3 Character Key Information Extraction

For the content of the form filling, there are two types: printed and handwritten. To clarify, we process the printed and handwritten forms separately to ensure accurate extraction of key information from both types of content filling.

3.1 Printed Information

When processing forms that contains only printed font, the form regions are first detected, and then the table structure of the archive image is recognized based on OpenCV. The image is segmented into sub-images according to the table structure. The outcomes of segmenting archival sub-images are showcased in Fig. 3.

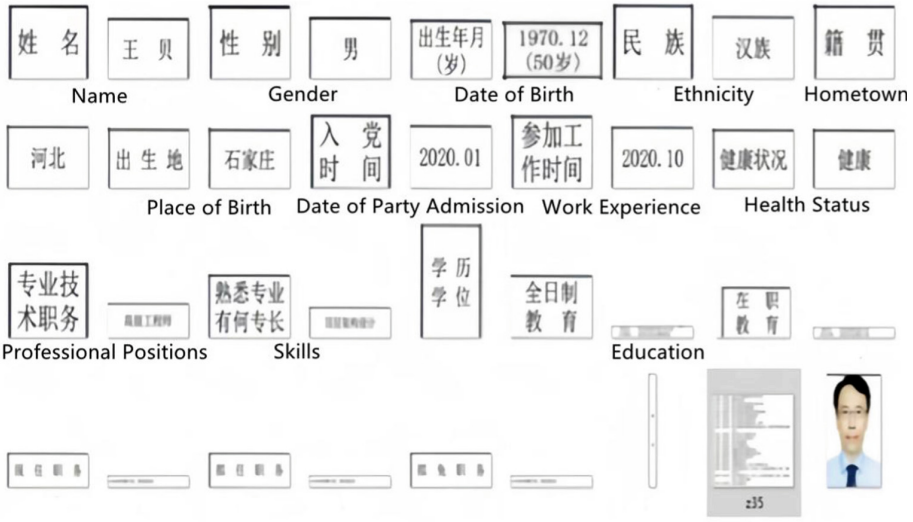

Fig. 3. The outcomes of segmenting archival sub-images.

Then the sub-images are input into a Chinese lightweight OCR model trained in advance to recognize and classify the text in the images using OCR technology [4]. The PP-OCR [5] technique is used to understand the recognized text content. The predicted results for each archive are written into a text file, converted, and finally stored as a CSV file that is easy for statistical analysis, achieving automatic extraction and structured processing of archive content. The structured extraction process is shown in the Fig. 4 below.

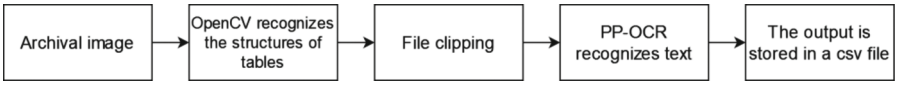

Fig. 4. The flow chart of structured extraction.

3.2 Handwritten Information

We propose a method that combines OCR with object detection. Before training the object detection model, OCR text recognition is used for auxiliary positioning and detection. The key printed words of the required fields are recognized to locate the range of the approval form [6, 10]. The entire archive image is then cropped into small images centered around the key information, reducing the feature extraction range during object detection. A personnel key information detection model based on the YOLOv5 network structure is then trained [7]. Through model inference prediction and image cropping functions. The tangible results of extracting pivotal personal information are showcased in Fig. 5.

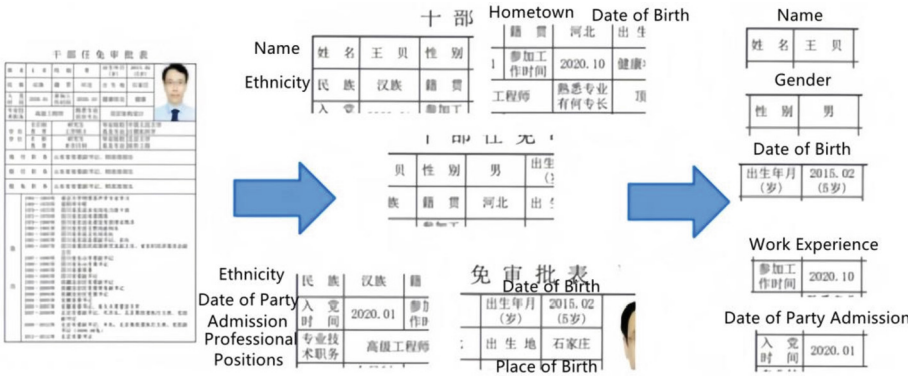

Fig. 5. The results of extracting pivotal personal information

And the name, gender, date of birth, important dates, and work start date in the archive are classified and output to the designated folder, and the information source is displayed. The workflow design is shown below Fig. 6.

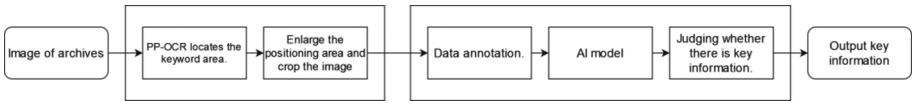

Fig. 6. ORC auxiliary detection technique and techniques for extracting key information

3.3 Ablation Experiment

We adopted the extraction strategy from archive ID photos and trained an archive key information extraction model using a lightweight object detection algorithm YOLOv5 to extract text blocks containing key information. To use it, the archive image to be processed is placed in a designated folder, and after model inference and cropping, all key elements in the archive can be obtained. All the key information is marked in the

tabular archive, and a single archive image usually contains three or more key pieces of information. The person key information detection and extraction model was trained on a real dataset of 399 complete archive images.

As shown in figure, the overall performance of the model on the real archive dataset is poor. Although the loss function of the training set and the test set shows a general downward trend, the loss function of the test set fluctuates sharply. Precision and recall are both low, the data is highly disparate, and the accuracy is also low. The mAP@0.5 reaches 0.85, and the MAP@0.5:0.95 is only 0.68. In summary, the person key information model has low comprehensive recognition accuracy and is not suitable for the person key information extraction module (Table 2).

Table 2. The detection effect on the real archive images data set.

	Images	Labels	P	R	mAP@.5	mAP@.5:.95
All labels	119	345	0.851	0.802	0.841	0.728
Name	119	132	0.844	0.822	0.894	0.765
Gender	119	93	0.923	0.903	0.955	0.814
Date fo birth	119	24	0.834	0.667	0.693	0.614
Employment date	119	12	0.823	0.833	0.814	0.726

4 Conclusion

In this paper, we discuss the technology of extracting key information from archival images of individuals and its domestic implementation, based on the problem scenarios of an actual project. We propose a method for extracting key content from archival images using deep learning. Specifically, we train a YOLOv5 object detection model to extract identity photos. We then use a combination of PP-OCR and YOLOv5 to extract key information such as names, gender, date of birth, and date of employment for content and handwriting comparison. We also perform structured extraction and recognition of fully printed personnel appointment and removal approval forms, storing the archival content as editable and searchable text in a CSV file. The archival image person information extraction technology can be used to batch identify and extract identity photos and key information from archival images. This method provides a way to digitize archival images, not only facilitating centralized comparison of identification photos of the same person during archival review, but also confirming the three ages and two histories of the archival review, verifying the key information and handwriting anti-counterfeiting, reducing the workload of review, and improving the efficiency of review.

Acknowledgements. We thank the ACs and reviewers. This work is partially supported by the Fundamental Research Funds for the Central Universities (3282023014), and the Project of Information Center of China North Industries Group Corporation Limited (20220100H0113).

References

1. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time Scene Text Detection with Differentiable Binarization. arXiv preprint arXiv:1911.08947
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
3. Nguyen, D.T., Nguyen, T.N., Kim, H., et al.: A high-throughput and power-efficient FPGA implementation of YOLO CNN for object detection. IEEE Trans. VLSI Syst. **27**(8), 1861–1873 (2019)
4. Sun, Y., Liu, J., Liu, W., Han, J., Ding, E., Liu, J.: Chinese street view text: Large-scale chinese text reading with partially supervised learning. In: Proceedings of the ICCV, pp. 9086–9095 (2019)
5. Du, Y., et al.: PP-OCR: A Practical Ultra Lightweight OCR System. arXiv preprint, arXiv:2009.09941
6. Liao, M., Wan, Z., Yao, C., et al.: Real-time scene text detection with differentiable binarization. Proc. AAAI Conf. Artif. Intell. **34**(07), 11474–11481 (2020)
7. Fischer, P., Smajic, A., Abrami, G., et al.: Multi-type-TD-TSR – extracting tables from document images using a multi-stage pipeline for table detection and table structure recognition: from OCR to structured table representations. In: Edelkamp, S., Möller, R., Rueckert, E. (eds.) KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, 27 Sep–1 Oct 2021, Proceedings, pp. 95–108. Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-87626-5_8
8. Zheng, Q., Zhu, J., Tang, H., et al.: Generalized label enhancement with sample correlations. IEEE Trans. Knowl. Data Eng. **35**(1), 482–495 (2021)
9. Wang, G., Xu, X., Shen, F., et al.: Cross-modal dynamic networks for video moment retrieval with text query. IEEE Trans. Multimedia **24**, 1221–1232 (2022)
10. Xu, X., Lu, H., Song, J., et al.: Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. IEEE Trans. Cybern. **50**(6), 2400–2413 (2019)