Applied Mathematics, Modeling and Computer Simulation
C. Chen (Ed.)

doi:10.3233/ATDE220079

An Open Digitization Tool for Extracting Scientific Curve Data in Portable Documents

Shichao ZHOU and Jun LU[1]

Commune of Scientific Engineers, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China

Abstract. Extracting original data in scientific figures embedded in documents precisely and efficiently remains a challenge, especially when it needs to be performed in high throughput way. To solve the automatic curve recognition problem, this paper proposes an integrated tool for extracting digital data from curve figures in portable documents, which mainly consists of picking/recognition/summary parts. During each extracting process, trained neural network firstly picks up figure pieces; the X-Y axes are then located by horizontal and vertical image projection and their labels are read using character recognition, which is followed by curve data recovery point by point; and finally the recognition result are summarized and sent back to the requester. This open tool is accessible and testable by anyone around the world via email, with open source on Github.

Keywords. Digitization of plotting graphs, figure curve recovery, scientific artificial intelligence.

1. Introduction

It is well known that artificial intelligence (AI) can greatly improve scientific research, which even include discovery of new superconducting materials [1] or writing a review book on lithium-ion batteries [2]. In comparison to recognition of characters, understanding of scientific figures is more challenging [3][4].

Among scientific figures, curve graphs are more difficult to be digitized and regenerated than others, because it need not only characters understanding of axial numbers and variables, but also quantitatively locating each point in plot- ted curves with correlation to numbers of scaling [5]. To digitize curve graphs, classical computer aid tools, such as Digitize-Pro or Origin software, usually users might be intimately involved during orientating points on curves or aligning axes. To perform curve recognition totally using AI, Bayesian network has been implemented, with accuracy about 63% [6]. Although Google and Baidu have already developed tools for recognition of complex

[1]Corresponding author, Jun LU, Commune of Scientific Engineers, Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China, https://orcid.org/0000-0003-3009-6683; Email: lujun@iphy.ac.cn.

graphs, it remains a problem that AI could hardly provide accurate and precise curve recognition service, which is efficient and totally automatic and easily accessible.

Therefore, to solve such problem, in this work we studied the tools for curve recognition and regeneration using step-by-step deep learning neural networks on a Python platform. The original input could be a portable document or image file, from which the AI-Curve tool picks each figure up and recovers original digital data successively. To improve the ease of use with appropriate security, the interactive interface has been implemented on email system for world-wide users without limitation of platform or software.

2. Method Demonstration

The main function of the AI-Curve tool was constructed simply step-by-step from portable documents files (pdf) to figure segments followed with plot digitization and report summary. As shown in figure 1, the input file format could be pdf or image file, where ”fitgz” library is called when necessary to convert from pdf to images.

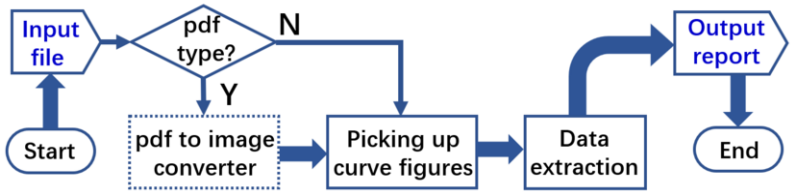

Figure 1. Principle flow chart of the open digitization tool for extracting scientific curves in portable documents.

As shown in figure 2, figure segments in the image page have been picked up using a deep learning neural network, where the YOLOv4 algorithm were used for training curve graph features [7]. The YOLOv4 algorithm core was used to train the model, prepare a large number of training data, and enhance the model by capturing image features and then generate the training model. The input pages or images to be recognized were detected and identified according to the training model data, and then the position and size of the figure segments could be obtained.

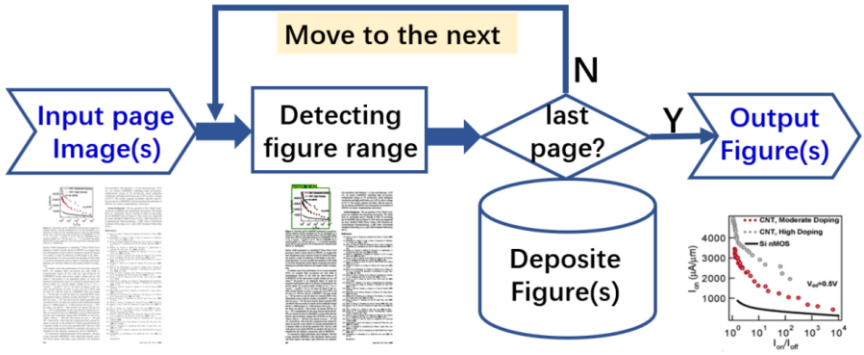

Figure 2. Demonstration process of picking up figure segments from each page of requested curve documents.

After that, the figure segments were processed to generate results with core curve data, as shown in figure 3, where coordinate scanning, label recognition, results summarizing, and output were executed successively. To reduce the im- age noise, figures were denoised by means of median filter and wavelet transform before main recognition procedures. After horizontal projection and vertical projection, the wave crest was used to determine the coordinate axis position. Then point by point scanning has been used in the coordinate axis to determine the curve in the plotting segment. After all curve data were digitized, curves were then regenerated, and the processing results were reported.

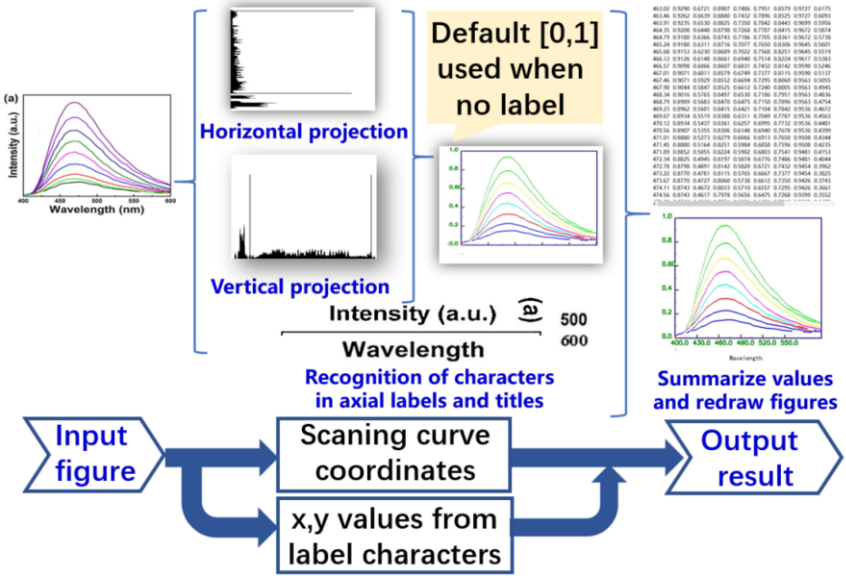

Figure 3. Core procedures of the open digitization tool include coordinate canning, label recognition, results summarizing and output.

3. Method Demonstration

3.1. Training Test

To train the deep learning AI model of the open digitization tool for curve figures, 806 pictures including graphs were selected from scientific publications, where the training and learning process adopted a Geforce GTX 1660 graphics card with memory of 6144 MB. The indicators training process were robustly convergent and shown in figure 4, including generalized intersection over union (GIoU) for measuring loss of bounding, objectness of decision, and accuracy. It is found that the recognition accuracy of the AI-Curve for picking figures from pictures is about 92% after 300 iterations.

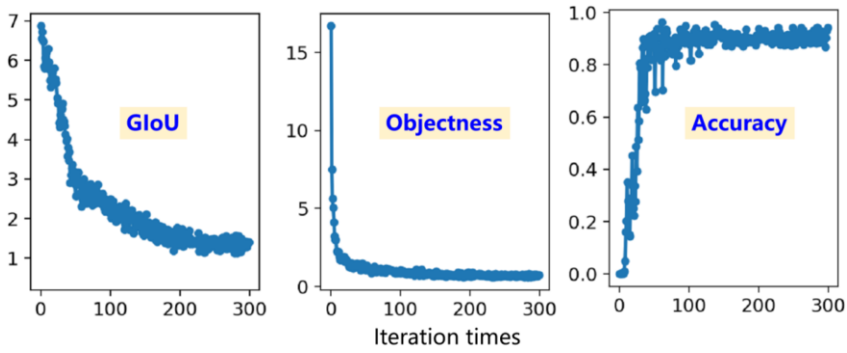

Figure 4. Training result of curve feature recognition using YOLOv4 algorithm.

3.2. Blurring Test

It is rather interesting and valuable for users to get the resistance to lower quality of input figures. As far as we concern, this performance has not been realized previously for AI related curve data extraction or regeneration. To obtain the capability of the AI-Curve, test curves has been degraded by Gaussian blur algorithm as introduced by Flusser et al. [8], while the absolute clarity of figures was evaluated by Laplacian sharpness index (LSI) calculated according to an automatic focus determination [9]. The test process and results have been shown in figure 5, where the blurring degree reach distance about 4 pixels, corresponding LSI about 1, the AI-Curve tool could not correctly recognize curve data anymore. It is also found that the recognition process could more easily regenerate the curve shape than recognize exact (x,y) data.

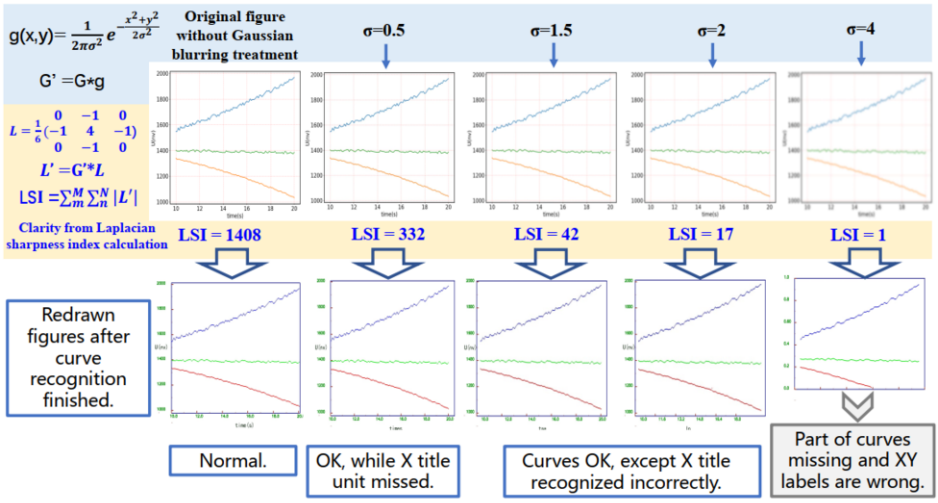

Figure 5. Blurring test procedures and results for the open digitization tool, showing the resistance to unclear figures.

3.3. Web Interface for Use

To improve the ease of use with appropriate security, the interactive interface has been implemented on email system for world-wide users without limitation of platform or software. The construction concept of the AI-Curve interface has been shown in figure 6, where every user around the world can test the open digitization tool through email to scienceai@sina.com with "curve" included in title and curve graphs attached in pdf or jpg file. The AI-Curve tool subsequently responses requests by checking mailbox, recognizing curves in attachment, summarizing a report, and then replying to the request per email. For those who are interesting in development of this tool, please review the source codes uploaded in Github [10].

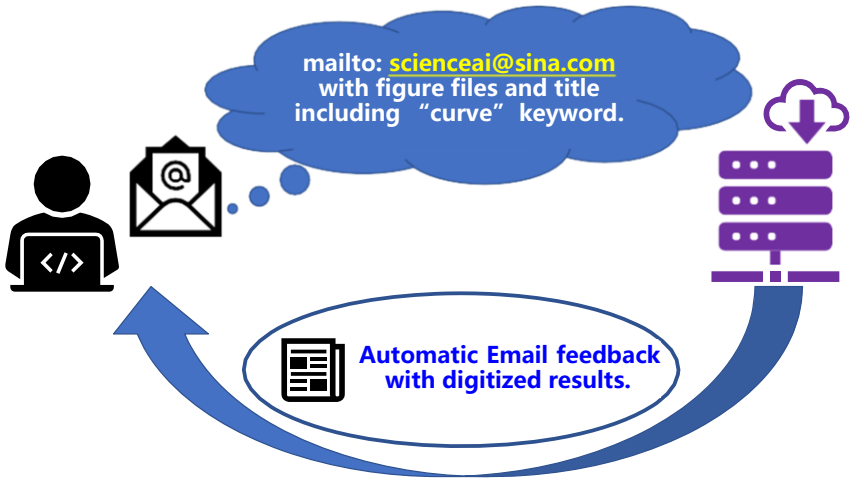

Figure 6. The framework of web interface for use of the open digitization tool.

4. Summary

This paper introduces a digitization tool for extracting scientific data from curve figures in portable documents. The tool consists of picking/recognition/summarizing parts for scientific curve and has been implemented by Python language. For clear document pages embedding curves, target figure region could be picked up with about 92% accuracy. When the original document includes unclear curve figures, the recognition process got more difficulties in recovery of exact (x,y) data than recognizing curve shape. This tool is independent on user platform or application environment and can be easily accessed by everyone with web mail interface. We greatly appreciate tests and feedback from world-wide users via email to scienceai@sina.com with "curve" included in email title and curve graphs attached in pdf or jpg file.

Acknowledgement

This work is under supports from National Natural Science Foundation of China (No. 51327806), Fujian Institute of Innovation and Youth Innovation Promotion Association of Chinese Academy of Sciences (No.2018009).

References

[1] Stanev V, Oses C, Kusne A, Rodriguez E, Paglione J, Curtarolo S, Takeuchi I 2018 Machine learning modeling of superconducting critical temperature, npj *Computational Materials* 4(1) https://doi.org/10.1038/s41524- 018-0085-8, http://www.nature.com/articles/s41524-018-0085-8

[2] Writer, Beta: Lithium-Ion Batteries: A Machine-Generated Summary of Current Research. Springer International Publishing (2019), https://www.springer.com/gp/book/9783030167998

[3] Liu Y, Lu X, Qin Y, Tang Z, Xu J 2013 Proc. SPIE 8654 Visualization and Data Analysis: Review of chart recognition in document images, https://doi.org/10.1117/12.2008467

[4] Llados J, Kwon Y B: Graphics Recognition: Recent Advances and Perspectives. Springer International Publishing (2003)

[5] Lu X, Kataria S, Brouwer W, Wang J, Mitra P, Giles C 2009 Automated analysis of images in documents for intelligent document search. International Journal on Document Analysis and Recognition, 12(2), 65–81. http://link.springer.com/10.1007/s10032-009-0081-0

[6] Nair R, Sankaran N, Nwogu I, Govindaraju V 2016 12th IAPR Workshop on Document Analysis Systems: Understanding line plots using bayesian network. http://ieeexplore.ieee.org/document/7490102/

[7] Bochkovskiy A, Wang C Y, Liao H M 2020 Yolov4: Optimal speed and accuracy of object detection (2020), http://arxiv.org/abs/2004.10934v1

[8] Flusser J, Farokhi S, Hoschl C, Suk T, Zitova B, Pedone M 2016 Recognition of images degraded by gaussian blur. IEEE Transactions on Image Processing, 25(2), 790–806. http://ieeexplore.ieee.org/document/7364266/

[9] Pech-Pacheco J, Cristobal G, Chamorro-Martinez, J, Fernandez-Valdivia J 15th International Conference on Pattern Recognition: Diatom autofocusing in brightfield microscopy: a comparative study. http://ieeexplore.ieee.org/document/903548/

[10] Open source link of the ScienceAI-Curve program: https://github.com/ZHOUSHCH/curve