

Automatic digitization of time series plots from published sources with EVA application

Veronika Voronova¹, Leonid Stolbov¹, Kirill Peskov^{1,2,3}, Kirill Zhudnikov^{1,2}, Victor Sokolov^{1,3}

1 - M&S Decisions, Dubai, United Arab Emirates, 2 - Research Center of Model-Informed Drug Development, Sechenov First Moscow State Medical University, Moscow, Russia, 3 - Marchuk Institute of Numerical Mathematics of the Russian Academy of Sciences (INM RAS), Moscow, Russia



Introduction

The extraction of quantitative data from published scientific plots is essential for generating datasets in meta-regression and quantitative systems pharmacology (QSP) modeling. Manual digitization remains a laborious and time-consuming bottleneck and its automation holds the potential to substantially accelerate pharmacometrics research.

Examples of digitization solution [1-5]

Automatic digitization

Raster graphics: Computer vision (CV)

Python libraries (e.g. ChartReader, ChartDete, LineEX)

Vector graphics: parsing

Large Language Models (LLM)-based approaches (Graphscribe).

Manual digitization

PlotDigitizer and Engauge Digitizer

Despite the variety of existing approaches, no ready-made solution for data digitization has been developed, as these approaches have the following limitations:

- CV and LLM usage is often constrained by the need for graphics processing unit (GPU) acceleration.
- CV and LLM algorithms can yield false or misleading results, and current libraries lack integrated tools for visual assessment and manual correction of the output and require coding skills for image processing.
- No algorithms has been proposed for analysis of pharmacometrics plots including time series data with error bars.

Objectives

We propose **EVA** – a programming interface and a part of **Simurg** ecosystem that automates data **E**xtraction, **V**isualization and **A**nalysis from pharmacometrics plots while allowing users to manually refine the results.

Methods

Data digitization workflow can be represented by a following diagram:

Coordinate system definition:

- Axes detection
- Labels and ticks detection
- Labels-ticks pairing
- Setup of regression models for pixel-value coordinate conversion

Data elements detection:

- Point detection and group assignment
- Error bar detection and error bar – point pairing

Point and error bar coordinates calculation

For the current version 1.0 of EVA software, we implemented this workflow to enable vector graphics processing implying parsing of .pdf files followed by extraction of the graphical object properties and rule-based chart element detection and classification to: (1) ticks and labels, (2) experimental points, (3) error bars as well as points grouping and labels-ticks and points-error bars pairing.

To facilitate the processing of raster plots, we plan to implement CV algorithms based on lightweight detection and segmentation models. As a first step, we have developed and tested a model for the detection of ticks and labels.

For coordinate conversion RANSAC regression models were utilized as done before [3], to minimize impact of outlying coordinates in case of incorrect label recognition. Automatic detection of logarithmic axes is performed by fitting RANSAC models under linear and exponential transformations, followed by model selection using an objective function.

The programming core was implemented in typescript and supplemented with a typescript-based interface enabling PDF document load and navigation, image selection and processing, visual check and manual correction of the digitization result, working with a final data table.

Results

Workflow

STEP 1
Upload .pdf to the web-based interface and crop region

STEP 2
Inspect elements detection and introduce necessary corrections

STEP 3
Run and inspect diagnostics

STEP 4
Save output in a csv format

Features

- .pdf files support
- Processing of time-series multigroup plots with error bars
- Automatic log-scale detection
- Possibility of manual result correction
- No dedicated GPU necessary

CV algorithms implementation

YOLOv10n model was trained and validated on PMC 2022 chart element detection data [https://github.com/pengyu965/ChartDete] using AdamW optimizer. Ticks and labels detections were followed by Tesseract optical character recognition.

Validation of the detection result

Metrics	Result
Mean Average Precision at 50% IoU	0.918
Precision	0.915
Recall	0.878

Conclusions

- A programming interface enabling automatic data extraction from vector graphics with an ability for manual correction of the result has been developed and will be adapted for other figure types and supplemented with CV algorithms for raster graphic processing.
- The proposed CV algorithm for automated ticks and labels detection shows high accuracy.

References

1. Wojtyniak JG, Britz H, Selzer D, Schwab M, Lehr T. CPT PSP . 2020 Jun;9(6):322-331. doi: 10.1002/psp4.12511. Epub 2020 Jun 16. PMID: 32543786; PMCID: PMC7306621.
2. Shivasankaran, V. P., Muhammad Yusuf Hassan, and Mayank Singh. 2023 IEEE/CVF (WACV). IEEE, 2023.
3. Yan, P., Ahmed, S., Doermann, D. (2023). ICDAR 2023. ICDAR 2023. Lecture Notes in Computer Science, vol 14187. Springer, Cham. https://doi.org/10.1007/978-3-031-41676-7_13
4. Z. Cheng, Q. Dai and A. G. Hauptmann, ICCV Paris, France, 2023, pp. 22145-22156, doi: 10.1109/ICCV51070.2023.02029.
5. https://github.com/pawankumar94/graphscribe-table-extractor

Poster S-10,
34th PAGE meeting
Dubrovnik, Croatia

Veronika.Voronova@msdecisions.tech

