FROM IMAGES TO SMART DATA: DIGITIZATION OF LOGISTIC DOCUMENTS

MOTIVATION

- Lithuania has about 50,000 active trucks each month, resulting in about 192 tonnes of documents annually.
- The manual entry into enterprise resource planning (ERP) systems is time-consuming and could consist of errors.

GOAL

- To create a framework for the digitization of logistic documents: Invoices, Receipts, CMRs
- To implement in the production workflow.

RESEARCH DESCRIPTION

1 Dataset

Total dataset consists of 50 GB of documents in PDF and JPG formats.

Initial dataset consists of 1500 manually selected documents (~500 for each class).

2 Data Preparation

Individual pages from PDF files were extracted and converted into JPG files.

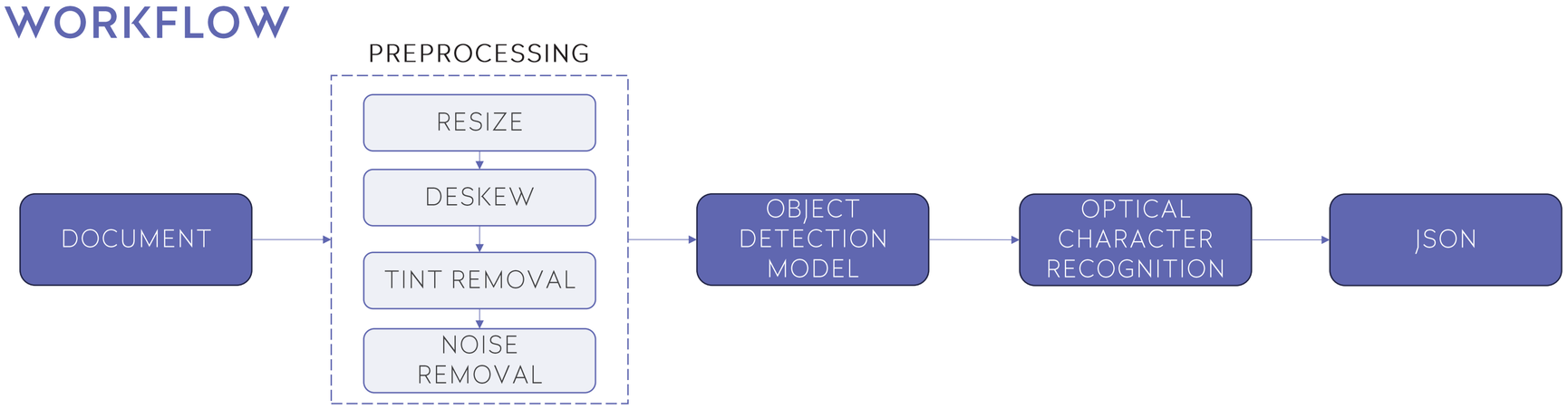
JPG files were preprocessed before annotation process.

3 Data Labeling

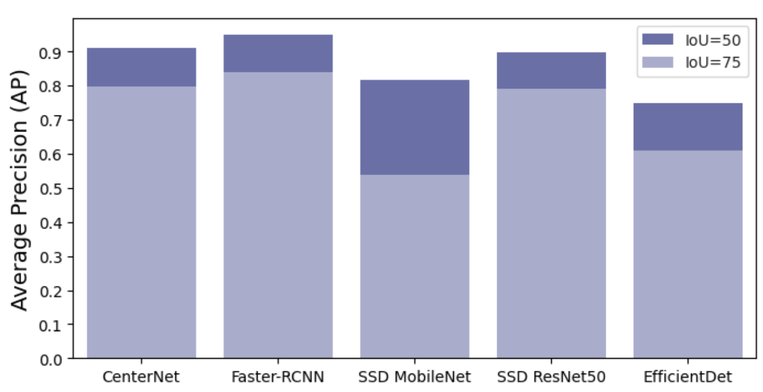
Initial dataset was labeled manually using LabelImg and saved in Pascal VOC format.

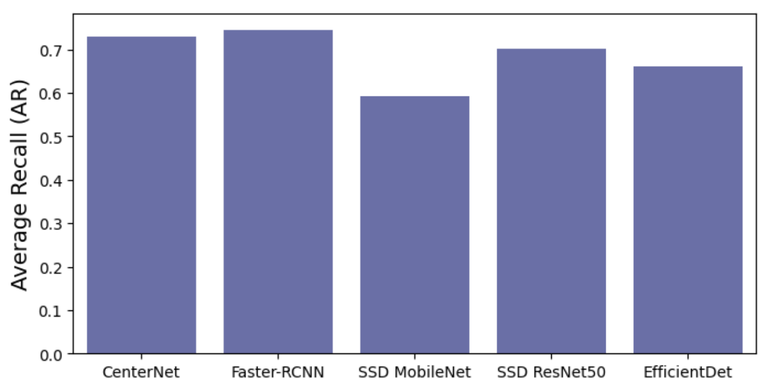
Full dataset will be labeled semi-automatically, utilizing a pre-trained object detection model and manual annotations adjustment.

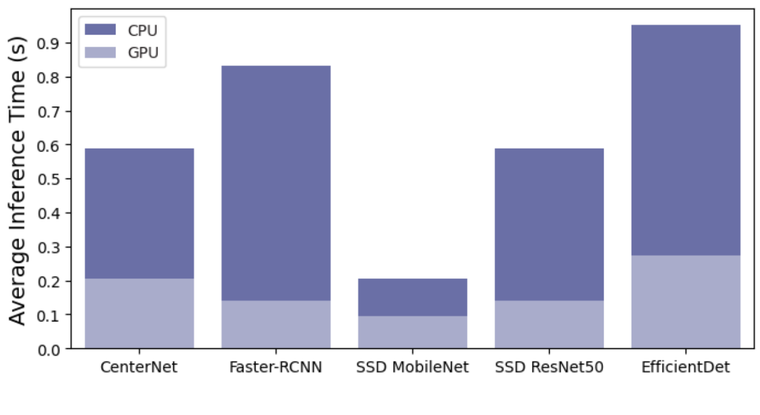
WORKFLOW



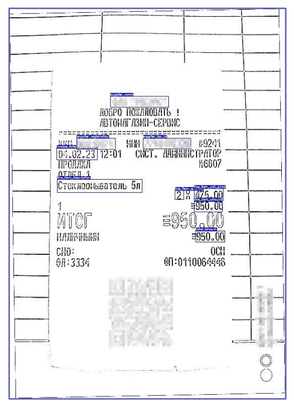
EVALUATIONS

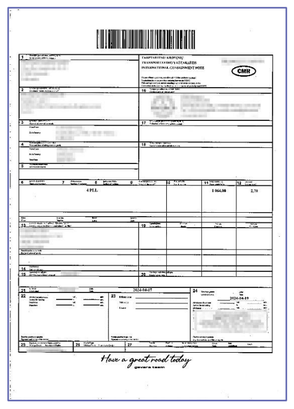


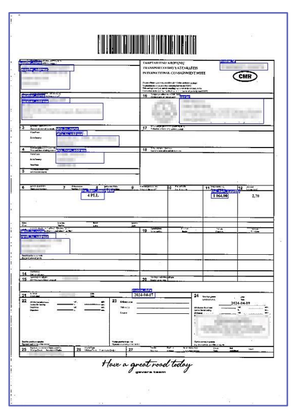


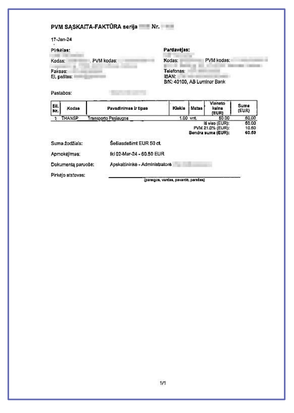


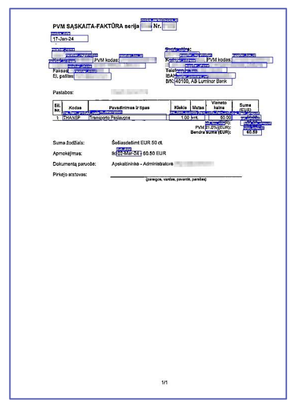
PREDICTIONS











CONCLUSIONS

- The Faster-RCNN model achieved an average precision (AP) of 0.95 at an Intersection over Union (IoU) threshold of 0.5, with an average recall (AR) of 0.75, within the same range.
- OCRs performances were manually assessed due to the lack of annotations, with Google OCR proving the best results.
- Current Faster-RCNN model will be used in automating a document labeling process.

AUTHORS:

Eimantas Zaranka
eimantas.zaranka@vdu.lt

Monika Zdanavičiūtė
monika.zdanaviciute@vdu.lt

Tomas Krilavičius
tomas.krilavicius@vdu.lt

Co-funded by the European Union