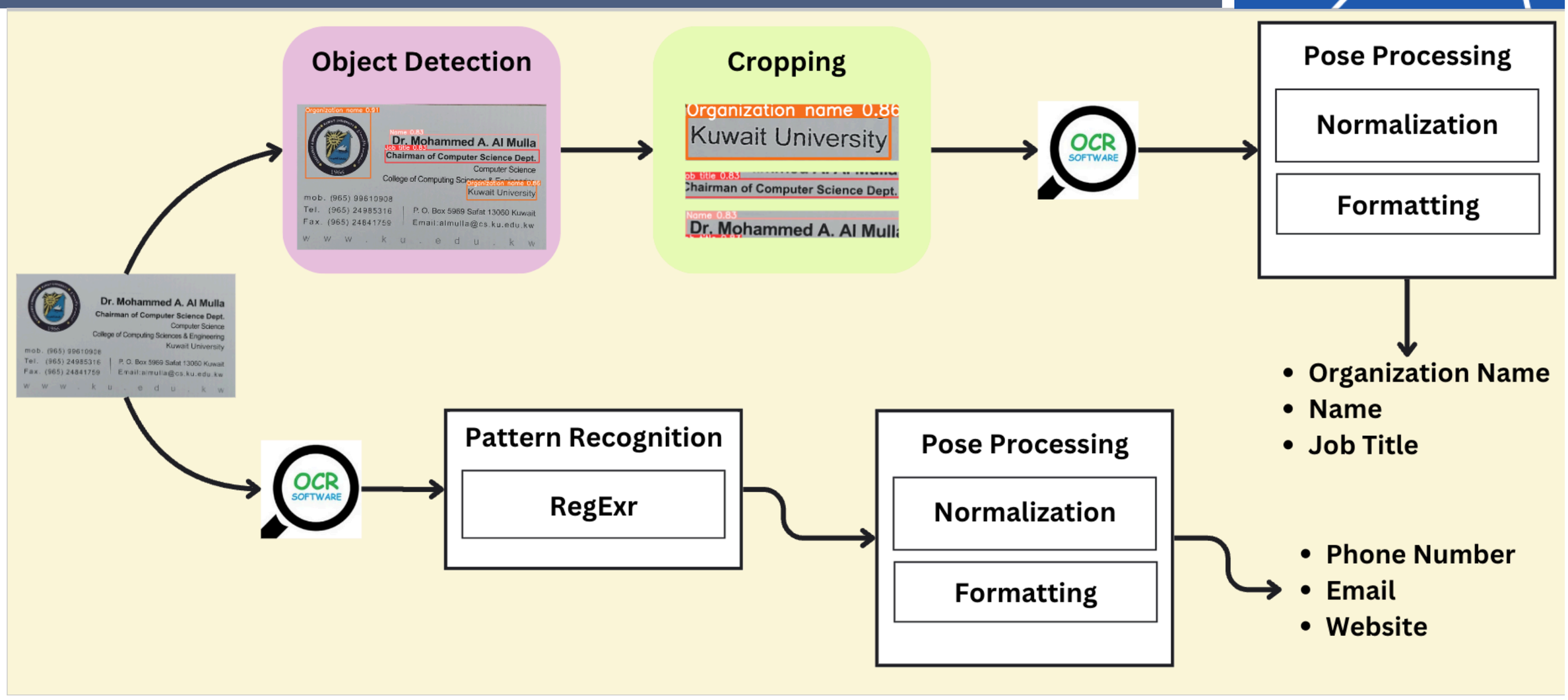
OCR PIPELINE SERVICE

Viphava Khlaisuwan (Ohm), Supervisor: Zein Zebib
CERN meyrin switzerland
Thai - CERN Collaboration Program under the initiative of H.R.H Princess Maha Chakri Sirindhorn
Software Engineering, Thammasat University, Thailand



INTRODUCTION

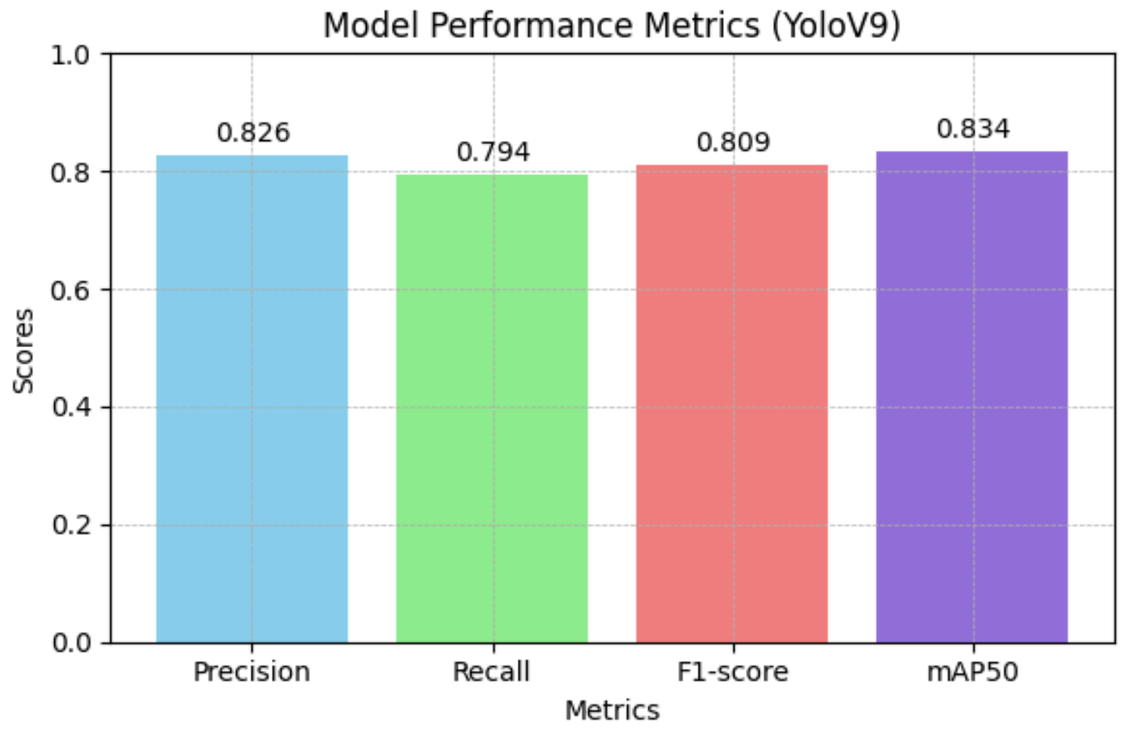
The Contack UP Application is designed to address the challenge faced by CERN members who frequently receive numerous new contact cards.

Manually entering contact information into digital formats is both time-consuming and prone to errors, especially given the diverse formats and international variations of these cards. To tackle this, we have developed an advanced OCR (Optical Character Recognition) pipeline that automates the extraction and entry of contact information. This pipeline accurately identifies and extracts key details such as names, organization names, job titles, phone numbers, emails, and websites, ensuring a seamless and efficient process for creating and managing contact lists.

The pipeline leverages object detection for precise localization of text regions, followed by OCR processing and pattern recognition to ensure high accuracy and reliability in diverse contexts.

OBJECTIVE

- Develop an advanced OCR pipeline to automate the extraction of contact information from business cards.
- Ensure accurate detection and extraction of names, organization names, and job titles using object detection.
- Implement robust OCR processing to handle diverse card formats and text variations.
- Utilize pattern recognition to accurately identify and extract phone numbers, emails, and websites.
- Normalize and format extracted data to maintain consistency and accuracy.
- Enhance productivity and reduce manual entry errors by automating the contact information management process.



Conclusion

The OCR pipeline for the Contack UP Application shows strong performance, with a precision of 0.826, recall of 0.794, F1-score of 0.809, and mAP50 of 0.834. These metrics indicate the pipeline's effectiveness in automating contact information extraction, though future improvements with a larger dataset and enhanced pattern recognition can further increase accuracy. This solution significantly enhances productivity and reduces manual effort for CERN members.

Limitations

- **Small dataset:** Only 274 contact card images were available for training the model.
- **Subjective labeling:** Labels were created based on personal judgment, which may affect the consistency and accuracy of the results.

Future Works

- Acquire a larger dataset and train the model to achieve higher accuracy.
- Enhance the pattern recognition of RegEx to handle a wider variety of formats.