Development of an On-Premise OCR Pipeline for Robust MRZ Extraction from Real-World Images of ID Documents

Constantin Jann Eberdorfer · Master Thesis, AI Engineering · FH Technikum Wien, 2026

1 Motivation & Problem

- Identity verification underpins telecom, banking and public services.
- The **Machine-Readable Zone (MRZ)** encodes identity data in a strict ICAO format with **checksums.** One misread can invalidate the whole record [1].
- MRZ extraction is **largely solved** for controlled **scans** [2].
- Real-world **mobile photos** remain **hard**: glare, blur, perspective and occlusion.
- Commercial solutions (e.g. AWS, Regula) are closed black boxes, expensive and a privacy risk.
- **No** production-ready, **open source system** for robust MRZ extraction exists [2].

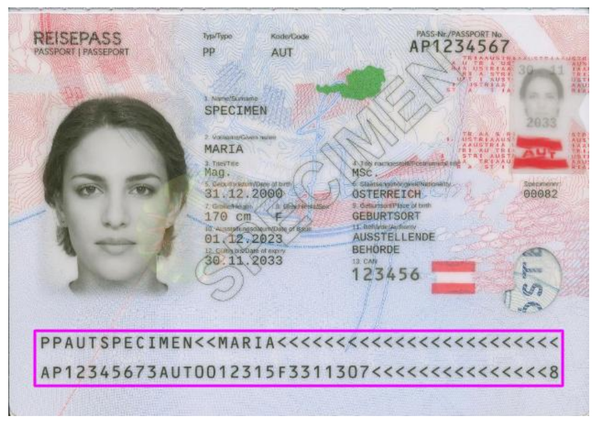

Figure 1: A **TD3 MRZ** (two lines, highlighted) on a specimen passport [3].

2 Proposed Method

OBJECTIVES

1. **MRZ Detection**: compare four segmentation backbones: HED-MRZ, U-Net, DINOv2-S, DINOv3-L.
2. **MRZ Optical Character Recognition (OCR)**: test whether MRZ-specific (SVTRv2, restricted to the ICAO alphabet) beats generic OCR (PaddleOCR v3/v4/v5, AWS Textract).
3. **End-to-end (E2E) Pipeline**: quantify the gap between isolated OCR on ground-truth crops and the E2E pipeline.

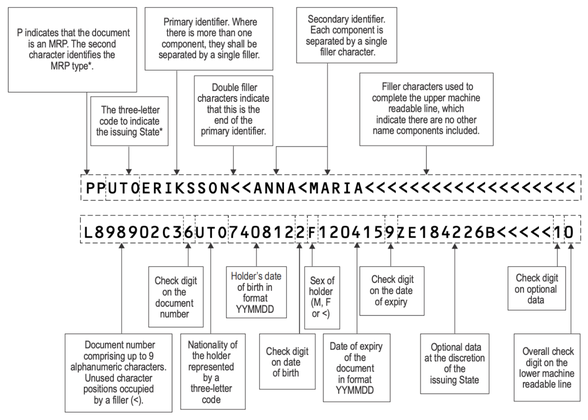

Figure 2: TD3 field layout and **check-digit** positions that make the MRZ self-verifying.

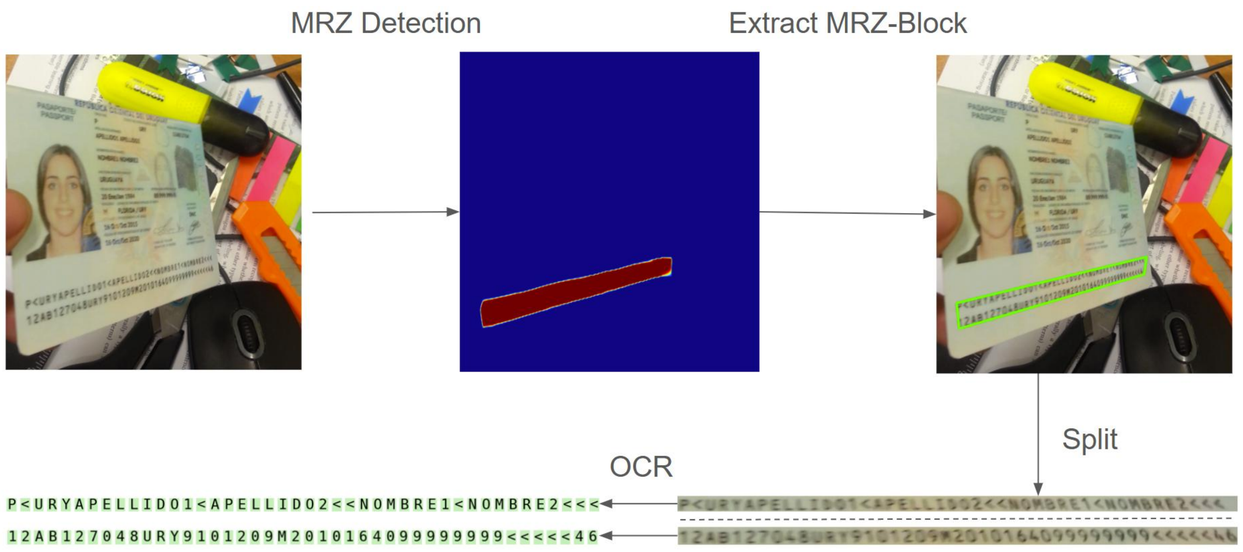

Figure 3: Overview of the E2E pipeline. Example: MIDV500 URY passport PS PS46_10

Two **learned stages** (MRZ detection, OCR) plus rule-based block extraction and line splitting, measured end-to-end under a **cross-dataset protocol** that avoids leakage from near-duplicate frames.

Trained & tested on public **MIDV-500 / 2020 / LAIT** plus a **private set** of 1,873 real smartphone captures (54 countries, TD1 + TD3); the recognizer is pretrained on synthetically generated MRZ lines.

3 Results MRZ OCR

98.6%

Best **PCR** on MIDV-500: **+5.6 pp** over prior SOTA

METHOD (Test Set MIDV500)	PCR %
2-step CNN by Chernyshova et al. [4] (prev. SOTA)	92.98
PaddleOCRv3 reported by Gayer et al. [2]	92.40
PaddleOCR v3 (reproduced)	91.73
Proposed: SVTRv2 (MRZ-adapted)	**98.62**

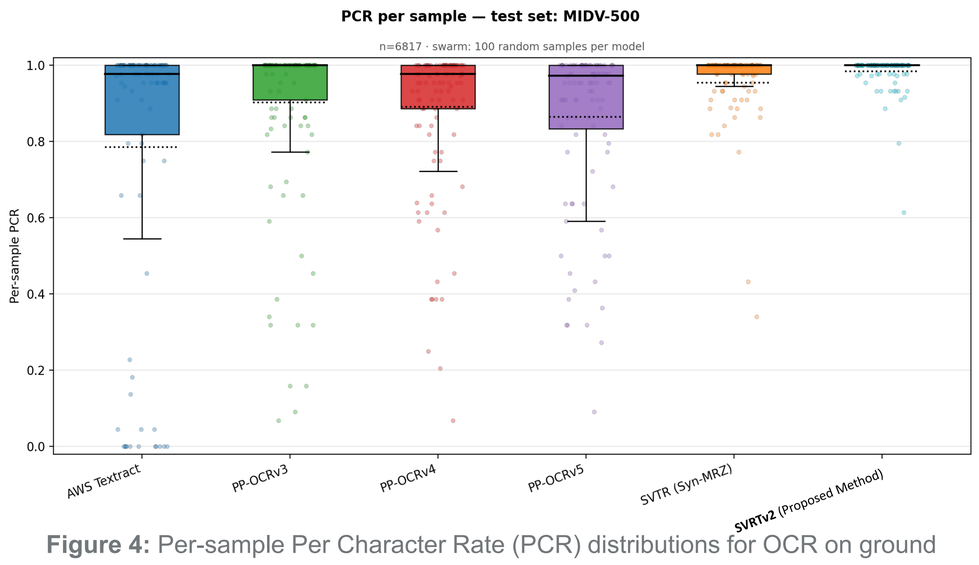

Figure 4: Per-sample Per Character Rate (PCR) distributions for OCR on ground truth crops.

4 Results E2E Pipeline

96.4%

Full-MRZ accuracy on the private test set.

Train Set	Test Set	Full MRZ Accuracy (%)	CER (%)
MIDV2020	MIDV500	70.78	1.79 ± 6.01
MIDV2020	MIDVLAIT	84.62	0.32 ± 1.25
MIDV500	MIDV2020	70.19	3.88 ± 11.23
MIDV500	MIDVLAIT	75.54	0.85 ± 3.87
Private	Private	**96.42**	**0.07 ± 0.41**

E2E pipeline performance with DINOv3-L and SVTRv2.

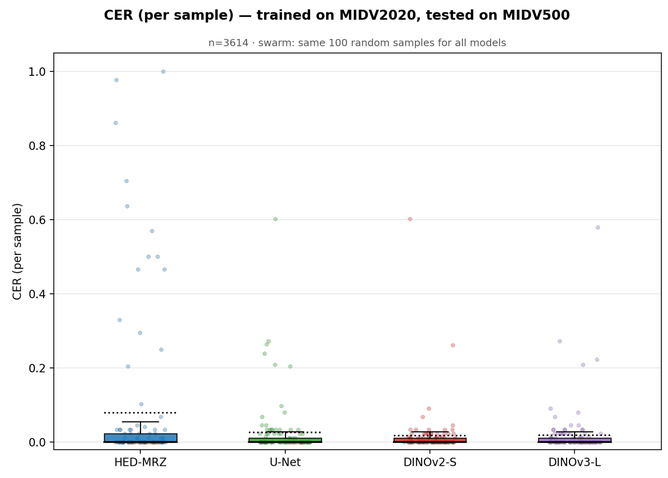

Figure 5: Per-sample CER distributions for the E2E pipeline

5 Discussion

KEY TAKEAWAYS

- A **robust on-premise MRZ pipeline is feasible** without closed commercial OCR.
- Recipe: **strong segmentation backbone + MRZ-adapted OCR** (synthetic pretraining → fine-tuning on real data).
- **Generic OCR does not transfer**; MRZ is a domain-specific OCR problem.
- Once detection is strong, **OCR is the dominant error**; leftovers are glyph confusions (S→5, Q→G, X→V).

LIMITATIONS & OUTLOOK

- Private set is **Swiss-dominated and lacks TD2**; broad generality not claimed.
- **Heuristic** rectification / line-splitting fails on strongly curved documents.
- Future work: **checksum-guided correction**, broader data and deployment metrics (latency, memory).

REFERENCES

[1] Machine Readable Travel Documents, Part 4: Machine Readable Passports (Doc 9303), Available: https://www.icao.int/publications/doc-series/doc-9303 [2] A. V. Gayer, Y. S. Chernyshova, and V. V. Arlazarov, "Recognition of machine-readablezone in identity documents: A review," IEEE Access [3] Wikimedia Commons contributors. "Passport of austria (2024) data page.". Available: https://commons.wikimedia.org/wiki/File:Passport_of_Austria_(2024)_data_page.jpg [4] Y. S. Chernyshova, A. V. Sheshkus, and V. V. Arlazarov, "Two-step cnn framework for textline recognition in camera-captured images," IEEE Access