Robust Semantic Digitization of Unconstrained Handwritten Medical Forms: A Hybrid Approach Using Deep Learning and Heuristic Layout Signatures

Nika Ramin, Omid Mohebi, Alireza Nourizadeh, Mahdi Tajdari

Abstract

Digitizing historical or operational paper forms remains a significant challenge, particularly for documents lacking machine-readable identifiers (QR/Barcodes) and written in Right-to-Left (RTL) scripts like Farsi. This paper presents an end-to-end automated pipeline for the extraction, validation, and semantic mapping of handwritten medical forms. We propose a novel **Page Recognition module** that utilizes heuristic layout signatures to identify document pages without explicit markers. Furthermore, we introduce a **Self-Correcting Table Extraction** method employing **DBSCAN clustering** to statistically remove ghost detections and recover missing cells. Experimental results on real-world medical forms demonstrate the system's ability to handle ambiguous marks (strikethroughs) and mixed alphanumeric handwriting, providing a robust solution for intelligent document processing (IDP).

1 Introduction

The digitization of medical records is critical for modern healthcare analytics, yet a vast quantity of data remains locked in physical paper forms. While Optical Character Recognition (OCR) has matured for printed text, processing handwritten forms in low-resource languages like Farsi (Persian) presents unique structural and semantic challenges.

1.1 Problem Statement

Current commercial IDP solutions often fail when applied to legacy medical forms due to two primary constraints:

- **Ambiguity in Optical Mark Recognition (OMR):** Standard pixel-counting algorithms struggle to distinguish between a valid selection and a "correction" (e.g., a user crossing out a box to select another). Recent studies highlight that while deep learning classifiers can categorize marks as "confirmed" or "crossed-out," they often require massive annotated datasets that are unavailable for specific institutional forms.
- **The "Farsi Gap":** Handwritten Farsi/Arabic recognition is complicated by the cursive nature of the script, where character shapes change based on position, and the presence of diacritics (dots) that are easily confused with noise. Furthermore, many legacy forms lack barcodes or QR codes, rendering standard template-matching algorithms ineffective when pages are scanned out of order.

1.2 Contribution

We propose a hybrid pipeline that integrates state-of-the-art (SOTA) object detection (RT-DETR) with geometric heuristics. Our system uniquely combines a **layout signature** approach for barcode-free page recognition with a **statistical error correction** module that mathematically "heals" imperfect model detections.