AI-Based Automated Customs Declaration Recognition and Intelligent Archiving System

Jerry Xing

A01354731

Introduction

Many companies still rely on manual transcription when processing import customs declarations—copying key information from scanned images into Excel for statistics, accounting, and financial management. This process is repetitive, time-consuming, and error-prone. Through communication with my family's enterprise, which handles financial operations, we confirmed that this workflow is highly automatable.

Therefore, this project aims to build an AI-powered document-understanding-based customs declaration recognition and intelligent archiving system. The goal is to locate individual documents in combined scans automatically, understand handwritten and printed content, extract key fields, and produce structured data—significantly reducing manual input costs and improving accuracy and efficiency.

Background and Problem Statement

Based on communication with my family members, in real business scenarios, customs declarations are often batch-scanned or photographed by freight forwarders, producing large images/PDFs containing multiple declarations. These documents include both printed and handwritten content, as well as issues such as overwriting, cursive writing, residual ink, scattered noise, blur, or tilt.

Currently, financial staff must manually transcribe all information into Excel, then follow a "manual entry → manual verification" dual process. Workload grows linearly, or worse, with document volume, while still encountering errors such as misread numbers, missing fields, or currency confusion.

Given this context, what enterprises need is not a simple OCR tool, but a complete end-to-end document AI pipeline:
One that can automatically segment multiple declarations from combined scans, remain robust against handwriting/noise, extract fields, produce structured results, and support Excel/report generation and downstream analysis.

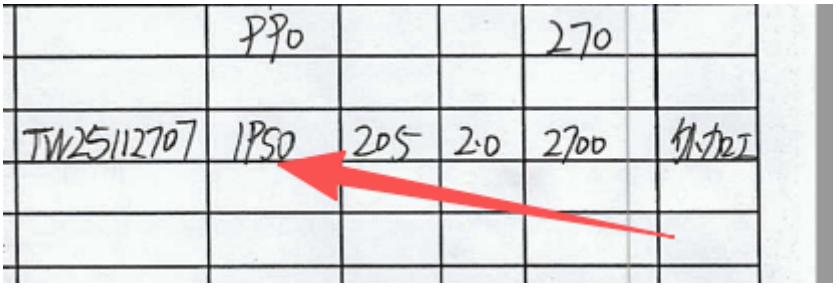
Technology Review and Existing Solutions

1. Capabilities and Limitations of Traditional OCR

Traditional OCR (e.g., Tesseract, printed-text PaddleOCR models, cloud OCR services) works well on clean printed text but faces inherent bottlenecks for this project's real-world data:

- **Cannot process multi-document segmentation in combined scans:** OCR "recognizes text," but does not "locate multiple declarations from one large image."

- **Poor robustness to handwritten/connected/overwritten content:** HTR (Handwritten Text Recognition) is a different domain from printed OCR.
- **Similar-looking characters:** Some handwriting habits commonly produce confusing pairs: "2" vs "z", "9" vs "q", "I (uppercase i)" vs "l (lowercase L)" vs "1". Heavily reliant on contextual reasoning. *(Page 2, Figure 1 shows a handwritten "9" easily mistaken as the letter "p")*



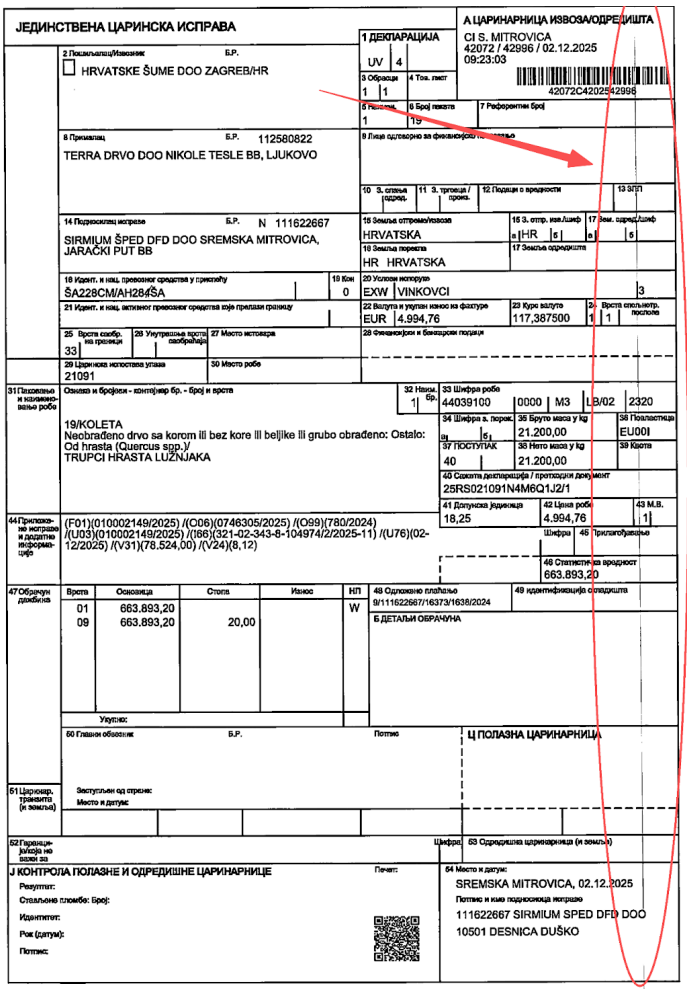
Page 2, Figure 1

- **Lack of semantic judgment for noise strokes:** OCR often mistakes non-text strokes/noise as characters.

Typical Failure Cases in Noisy Documents

- **Dry-ink test strokes:** Writers often scribble lines to check ink flow; traditional OCR misreads them as characters: `11111`, `/////`, `---`, `一一`.

- **Ink leakage/smudges / line-shaped noise:** Sparse spots may be misread as punctuation or vertical bars. *(Page 3, Figure 2 shows a long vertical noise line caused by scanning artifacts.)*



ЈЕДИНСТВЕНА ЦАРИНСКА ИСПРАВА

A ЦАРИНАРНИЦА ИЗВОЗА/ОДРЕДИШТА
CI S. MITROVICA
42072 / 42986 / 02.12.2025
09:23:03
42072C420254299B

1 ДЕКЛАРАЦИЈА: UV 4

2 Пошиљалац/Извозник: HRVATSKE ŠUME DOO ZAGREB/HR

8 Прималац: TERRA DRVO DOO NIKOLE TESLE BB, LJUKOVO

14 Подносилац исправе: SIRMIUM ŠPED DFD DOO SREMSKA MITROVICA, JARAČKI PUT BB

15 Земља отпреме/извоза: HRVATSKA

16 Земља порекла: HR HRVATSKA

17 Земља одредишта:

18 Идент. и нац. превозног средства у приколици: ŠA228CM/AH284ŠA

20 Услови испоруке: EXW VINKOVCI

22 Валута и укупан износ по фактури: EUR 4.994,76

23 Курс валуте: 117,387500

25 Врста саобр. на граници: 33

29 Царинска испостава утовара: 21091

31 Паковање и количина робе: 19/KOLETA
Neobrađeno drvo sa korom ili bez kore ili beljike ili grubo obrađeno: Ostalo:
Od hrasta (Quercus spp.)/
TRUPCI HRASTA LUŽNJAKA

32 Ниск. бр.: 1

33 Шифра робе: 44039100 | 0000 | M3 | LB/02 | 2320

34 Шифра з. порек.:

35 Бруто маса у kg: 21.200,00

36 Повлашћења: EU001

37 ПОСТУПАК: 40

38 Нето маса у kg: 21.200,00

39 Квота:

40 Сродна декларација / претходни документ: 25RS021091N4M6Q1J2/1

41 Допунска јединица: 18,25

42 Цена робе: 4.994,76

43 М.В.: 1

44 Прилагано исправе и додатне информације: (F01)(010002149/2025) /(C06)(0746305/2025) /(O99)(780/2024) /(U03)(010002149/2025) /(I66)(321-02-343-8-104974/2/2025-11) /(U76)(02-12/2025) /(V31)(78.524,00) /(V24)(8,12)

Шифра:

46 Статистичка вредност: 663.893,20

47 Обрачун дажбина

Врста	Основица	Стопа	Износ	НП
01	663.893,20			W
09	663.893,20	20,00		
Укупно:				

48 Одложено плаћање: 9/111622667/16373/1638/2024

49 Идентификација складишта:

6 ДЕТАЉИ ОБРАЧУНА

60 Главни обвезник: Б.Р. Потпис:

Ц ПОЛАЗНА ЦАРИНАРНИЦА

61 Царинар, траженто (и земља): Заступљен од стране: Место и датум:

52 Гаранција (изузета на основу):

Шифра: 53 Одредишна царинарница (и земља):

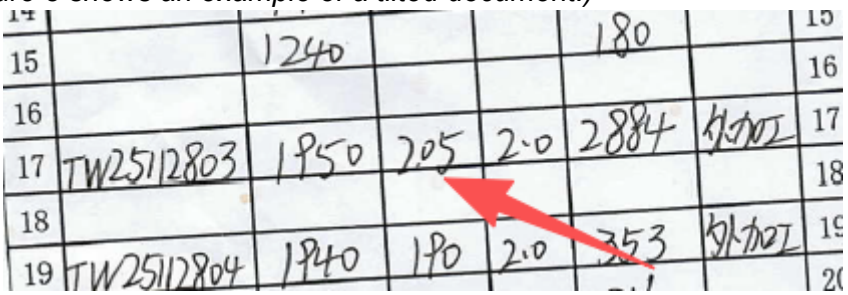
J КОНТРОЛА ПОЛАЗНЕ И ОДРЕДИШНЕ ЦАРИНАРНИЦЕ
Реквизит:
Стављене пломбе: Број:
Идентитет:
Рок (датум):
Потпис:
Печат:

64 Место и датум: SREMSKA MITROVICA, 02.12.2025
Потпис и име подносиоца исправе
111622667 SIRMIUM SPED DFD DOO
10501 DESNICA ĐUŠKO

Page 3, Figure 2

- **Ambiguous dots:** Users may dot near fields, or parts of digits may look like a dot, causing decimal/period errors (often disastrous). *(Page 3, Figure 3 shows the number “205” appearing like “2.05” due to table obstruction.)*

- **Tilted scans:** Slanted pages degrade recognition quality. *(Page 3, Figure 3 shows an example of a tilted document.)*



Page 3, Figure 3

- **Overwriting:** Modifications may resemble another character (e.g., a slashed “0” that looks like “8”).

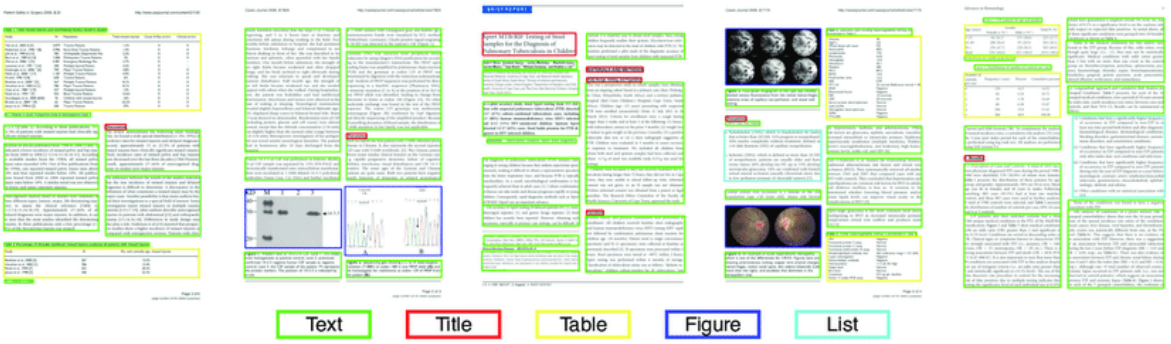
Conclusion

Under the conditions of “combined scans + handwriting + noisy documents,” traditional OCR **alone** is neither viable nor reliable.

2. Document Instance Detection and Layout Analysis

To extract each customs declaration from a combined scan, page layout and instance detection techniques are required:

- **Instance detection/segmentation:** YOLO models can detect bounding boxes of declarations and crop them out.
- **Layout analysis:** LayoutParser can identify table lines, cell positions, field areas, and content regions. *(Page 3, Figure 4 shows an example of layout recognition.)*



Page 3, Figure 4

3. Handwritten Text Recognition (HTR) and Semantic Validation

- **HTR models:** Handwriting-oriented PaddleOCR models, capable of handling connected writing and stylistic variations.
- **Multimodal Large Models (VLMs):** Vision-capable general models (e.g., GPT5.1) can combine context to perform semantic validation—e.g., distinguishing noise strokes from real characters.

4. Key Information Extraction (KIE) and Table Mapping

- LayoutLMv3 can combine text and spatial layout to perform KIE, reconstruct table structures, and map them to a defined schema.

Proposed Solution

Overall Concept

Multi-document segmentation → Handwritten/printed mixed recognition → Field extraction → Structured validation → Export

Processing Workflow

1. **Batch Input and Preprocessing**
   - Support uploading combined scanned images/PDF.
   - Perform de-skewing, de-noising, and contrast enhancement.
   - Rough document region segmentation; remove blank borders.
2. **Document Detection and Instance Segmentation**
   - Use YOLO to detect boundaries of each customs declaration and crop them out.

3. **Text and Content Recognition**
    - **Router:** determines whether the region is printed/handwritten/mixed.
    - **HTR branch**: PaddleOCR for handwritten content.
    - **Printed-text branch**: Traditional OCR for printed text.
4. **Key Field Extraction (KIE) and Semantic Validation**
    - LayoutLMv3 outputs structured JSON.
    - **Semantic and constraint checks:**
        - Number/quantity/currency/date type & range validation
        - Currency allowlist (CNY/USD/EUR/etc.)
        - Date normalization
        - Contextual correction of decimal/punctuation anomalies
    - **Noise suppression:** Using VLM (e.g., removing dry-ink lines, dots, overwriting artifacts)
5. **Structured Output and Back-Filling**
    - Map results to Excel.
    - Produce "one combined scan → multiple Excel records."
    - Save cropped local patches as evidence for human verification.
6. **Human-AI Collaboration**
    - Review interface showing values + image regions.
    - One-click correction and active learning: corrected samples fed back for continuous improvement.

Expected Outcomes

- Automatically detect and split each declaration from combined scanned images containing multiple documents.
- Achieve ~80% accuracy for key-field extraction.
- Reduce manual entry time by at least 60%.
- Maintain robustness under noisy document conditions: dry-ink strokes, smudges, tilt, and overwriting.
- Support exporting standardized Excel outputs for downstream financial analysis.

Minimum Viable Product

- **Input:** A combined scan containing 5–20 customs declarations.
- **Output:** One structured record (Excel/CSV/JSON) per declaration.
- **Must satisfy outcomes described above.**

Technologies Used

- **Backend & workflow orchestration:** Java + Spring Boot
- **Document detection & segmentation:** Python + PyTorch (YOLO)
- **Handwriting recognition (HTR):** PaddleOCR
- **KIE & rule-based validation:** LayoutLMv3 + regex + business rules
- **Data export:** Excel, JSON
- **Frontend review tool:** Web-based verification & correction interface