

Extracción automática de datos de facturas mediante reconocimiento óptico de caracteres y aprendizaje automático

Andrés Toledo¹; Luis Derico²; Susana Morales³; Aura Villalba⁴; Jorge Sanabria⁵; Alice Colmán⁶.

Ciencias Físico-Matemáticas e Ingenierías, Universidad Católica Nuestra Señora de la Asunción Campus Itapúa, Facultad de Ciencias y Tecnología.

Autor para correspondencia: abelardotoledo76@gmail.com

Introducción

Las empresas, tanto físicas como jurídicas, generan un alto volumen de facturas como resultado de sus actividades comerciales. Estos documentos presentan una gran diversidad en su estructura y formato, lo que suele requerir un procesamiento manual intensivo. La extracción manual y el registro de datos provenientes de las facturas implica un esfuerzo considerable de tiempo y recursos humanos, además de ser tareas repetitivas y propensas a errores, afectando directamente la eficiencia operativa y la precisión de los registros contables y financieros.

En este proyecto se investiga y desarrolla un *software* capaz de extraer automáticamente, con mínima intervención humana, los datos clave contenidos en diferentes tipos de facturas. Para lograrlo, se aplican técnicas de aprendizaje automático, lo que permite mejorar significativamente la precisión y la velocidad del proceso. Este trabajo contribuye al conocimiento sobre las tendencias tecnológicas actuales y ofrece una solución innovadora que optimiza tareas administrativas, promoviendo así una mayor eficiencia en la gestión empresarial.

Objetivo

Extraer automáticamente datos de facturas mediante Reconocimiento Óptico de Caracteres (OCR) y Aprendizaje Automático.

Metodología

Se aplicó un enfoque cuantitativo y una metodología de prototipado rápido. La investigación fue de tipo documental, aplicada y experimental:

Documental, para fundamentar teóricamente conceptos de OCR, IA y aprendizaje automático.

Aplicada y experimental, para desarrollar y probar un prototipo funcional de software. Se evaluaron motores OCR (PyTesseract y MMOCR), y modelos de clasificación (LayoutLM, BERT, SDMGR, YOLO, KNN).

Se recolectaron imágenes de facturas reales, se etiquetaron con LabelStudio, y se entrenaron modelos usando Python, Flask, OpenCV y Scikit-Learn.

Resultados

- El modelo MMOCR superó ampliamente a PyTesseract en precisión y adaptabilidad.
- El modelo SDMGR, integrado con MMOCR, combina información textual y visual, logrando una mejor clasificación de campos de facturas pero depende mucho de la calidad de la captura.
- El modelo YOLO, integrado con MMOCR, logró localizar con precisión y rapidez las áreas relevantes dentro de cada factura al centrarse únicamente en las regiones de interés.
- Con el uso de KNN se alcanza más rapidez en la detección de texto pero una pérdida de precisión en la clasificación de las categorías cuando los textos tienen similitudes estructurales, como las numéricas.

Conclusión

Los resultados evidencian que la integración de técnicas de OCR con aprendizaje automático permite automatizar con éxito la extracción de datos de facturas impresas.

Esta automatización reduce costos operativos, mejora la precisión de los registros contables y aumenta la eficiencia administrativa. Se concluye que el uso de herramientas como MMOCR y YOLO representa una alternativa viable para modernizar procesos empresariales.

Palabras Clave: Inteligencia Artificial, Aprendizaje Automático, Visión por Computadora, OCR.