COIMBRA BUSINESS SCHOOL

iscac Politécnico de Coimbra

Artificial Intelligence in Invoice Recognition: A Systematic Literature Review

Oleksandra Kukharska

ISCAC | 2023

COIMBRA BUSINESS SCHOOL

iscac Politécnico de Coimbra

Artificial Intelligence in Invoice Recognition: A Systematic Literature Review

Oleksandra Kukharska

Coimbra, October 2023

Oleksandra Kukharska

Artificial Intelligence in Invoice Recognition: A Systematic Literature Review

Dissertation requested from Coimbra Business School to fulfil the requirements for a Master's Degree in **Data Analysis and Decision Support Systems**, under the guidance of Professor António Rui Trigo Ribeiro.

Coimbra, October 2023

STATEMENT OF RESPONSABILITY

I declare to be the author of this dissertation, which constitutes an original and unpublished work, which has never been submitted to another higher education institute to obtain an academic degree or other qualification. I further certify that all citations are duly identified and that i am aware that plagiarism constitutes a serious lack of ethics, which may result in the cancellation of this dissertation.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

ACKNOWLEDGEMENTS

I want to express my gratitude to my professor, António Trigo, whose guidance, knowledge, and dedication have been invaluable in shaping this work and contributing to my academic growth. Thank you for your unwavering support and mentorship.

I would also like to extend my heartfelt appreciation to my family, whose unwavering support and encouragement have been my pillars of strength throughout this academic journey.

My friends have bought joy and balance my life during the challenges of this thesis, and I am truly grateful for their companionship.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

RESUMO

Numa era caracterizada por uma economia em crescimento e avanços rápidos na tecnologia da informação, a proliferação de dados de faturas acentua a necessidade urgente do reconhecimento automatizado de faturas. Os métodos manuais tradicionais, que há muito tempo têm sido usados para esta tarefa, demonstraram ser ineficazes, suscetíveis a erros e incapazes de lidar com o crescente volume de faturas. Esta investigação procura responder à necessidade de automatizar o reconhecimento de faturas, explorando, avaliando e avançando em algoritmos, técnicas e métodos de ponta no campo da Inteligência Artificial (IA). Esta investigação realiza uma Revisão Sistemática da Literatura abrangente para investigar as abordagens da Visão por Computador (CV, do inglês Computer Vision), englobando o pré-processamento de imagens, a Análise de Layouts (LA, do inglês Layout Analysis), o Reconhecimento Ótico de Caracteres (OCR, do inglês Optical Character Recognition) e a Extração de Informação (IE, do inglês Information Extraction). O objetivo é fornecer informações relevantes sobre esses componentes essenciais do reconhecimento de faturas, destacando a sua importância na busca de precisão e eficiência. Esta exploração visa contribuir para o desenvolvimento de sistemas automatizados mais eficazes na extração de informações de faturas, enfrentando os desafios apresentados por vários formatos e conteúdos. Os resultados indicam que, para a LA, a combinação de Mask Region-based Convolutional Neural Networks (M-RCNN) e Feature Pyramid Network (FPN) alcança resultados muito bons. No OCR, algoritmos como a Convolutional Recurrent Neural Network (CRNN), You Only Look Once version 4 (YOLOv4) e modelos inspirados na M-RCNN e Faster Region-based Convolutional Neural Network (F-RCNN) com ResNetXt-101 como espinha dorsal demonstram um desempenho notável. No que respeita à IE, os algoritmos inspirados na F-RCNN e na Region Proposal Network (RPN), Grid Convolutional Neural Network (G-CNN) e Layer Graph Convolutional Networks (LGCN), e Gated Graph Convolutional Network (GatedGCN) apresentam consistentemente os melhores resultados.

Palavras-chave: Fatura, Reconhecimento de Faturas, Inteligência Artificial, , Algoritmos, Visão por Computador, Extração de Dados

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

ABSTRACT

In the era marked by a flourishing economy and rapid advancements in information technology, the proliferation of invoice data has accentuated the urgent need for automated invoice recognition. Traditional manual methods, long relied upon for this task, have proven to be inefficient, error-prone, and incapable of coping with the rising volume of invoices. This research endeavours to addresses the imperative of automating invoice recognition by exploring, assessing, and advancing cutting-edge algorithms, techniques, and methods within the domain of Artificial Intelligence (AI).

This research conducts a comprehensive Systematic Literature Review (SLR) to investigate Computer Vision (CV) approaches, encompassing image preprocessing, Layout Analysis (LA), Optical Character Recognition (OCR), and Information Extraction (IE). The objective is to provide valuable insights into these fundamental components of invoice recognition, emphasizing their significance in achieving accuracy and efficiency. This exploration aims to contribute to the development of more effective automated systems for extracting information from invoices, addressing the challenges posed by diverse formats and content.

The results indicate that in LA, the combination of Mask Region-based Convolutional Neural Networks (M-RCNN) and Feature Pyramid Network (FPN) achieves goods results. In OCR, algorithms like Convolutional Recurrent Neural Network (CRNN), You Only Look Once version 4 (YOLOv4) and models inspired by M-RCNN and Faster Region-based Convolutional Neural Network (F-RCNN) with ResNetXt-101 as the backbone demonstrate remarkable performance. When it comes to IE, algorithms inspired by F-RCNN and Region Proposal Network (RPN), Grid Convolutional Neural Network (G-CNN) and Layer Graph Convolutional Networks (LGCN), and Gated Graph Convolutional Network (GatedGCN) consistently deliver the best results.

Keywords: Invoice, Invoice Recognition, Artificial Intelligence, Algorithms, Computer Vision, Data Extraction.

INDEX

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

INDEX OF FIGURES

INDEX OF TABLES

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

LIST OF ABBREVIATIONS AND ACRONYMS

AI – Artificial Intelligence

AP – Average Precision

ARRS – Automatic Receipt Recognition System

ASPP – Atrous Spatical Pyramid Pooling

BERT – Bidirectional Encoder Representations from Transformers

BiLSTM – Bidirectional Long Short-Term Memory

B-on – Online Knowledge Library

CCD – Charge Coupled Device

cGAN – Conditional Generative Adversarial Network

CNN – Convolutional Neural Network

Cond NN – Condensed Nearest Neighbor

CONV – Convolutional Layer

CPU – Central Process Unit

CRAFT - Character Region Awareness for Text Detection

CRNN – Convolutional Recurrent Neural Network

CUTIE – Convolutional Universal Text Information Extraction

CV – Computer Vision

DBSCAN - Desteny-Based Spatial Clustering of Applications with Noise

DL – Deep Learning

DFL-CNN – Double Focal Loss - Convolutional Neural Network

DN – Dense Net

DNN – Deep Neural Network

EAST – Efficient and Accurate Scene Text Detector

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

F-RCNN – Faster Region-based Convolutional Neural Network

FC – Fully Connected

FCN – Fully Convolutional Network

FPN – Feature Pyramid Network

FTCRF – Financial Ticket Character Recognition Framework

FTFDNet – Financial Ticket Faster Detection Network

GatedGCN – Gated Graph Convolutional Network

GCN – Graph Convolutional Networks

GNN – Graph Neural Network

GPU – Graphic Processing Unit

G-CNN – Grid Convolutional Neural Networks

GRU – Gated Recurrent Units

GUI – Graphical User Interface

ID - Identification

IE – Information Extraction

IBAN – International Bank Account Number

IoU – Intersection Over Union

KBKIDR – Knowledge-based Key Information Detection and Recognition method

LA – Layout Analysis

LGCN – Layer Graph Convolutional Networks

LSTM – Long Short-Term Memory

mAP – mean Average Precision

M-RCNN – Mask Region-based Convolutional Neural Networks

MS-RCNN – Mask Scoring Region-based Convolutional Neural Networks

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

MCNN – Modified Condensed Nearest Neighbor

MLP – Multi-Layer Perceptron

MSER – Maximally Stable Extremal Regions

NER – Named-Entity Recognition

NLP – Natural Language Processing

NMS – Non-Maximum Suppression

OCNN – Other Class Nearest Neighbor

OCR – Optical Character Recognition

PAN – Pixel Aggregation Network

PDF – Portable Document Format

PRISMA – Preferred Reporting Items for Systematic Review and Meta-Analyses

PSNR – Peak Signal-to-Noise Ratio

RAM – Random-Access Memory

RQ – Research Questions

RNN – Recurrent Neural Network

RPN – Region Proposal Network

SDAE – Stacked Denoising Autoencoder

SGD – Stochastic Gradient Descent

SIMD – Single Instruction Multiple Data

SLR – Systematic Literature Review

SNR – Signal-to-Noise Ratio

SoftAP – soft Average Precision

SSD – Single-Shot MultiBox Detector

SSIM – Structural Similarity Index

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

TIN – Tax Identification Number

UBL – Universal Business Language

UQI – Universal Image Quality Index

VAT – Value Added Tax

XML – Extensible Markup Language

YOLOv3 – You Only Look Once version 3

YOLOv4 – You Only Look Once version 4

YOLOv4-s – You Only Look Once version 4-small

YOLOv5 – You Only Look Once version 5

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: presidencia@iscac.pt; www.iscac.pt

INTRODUCTION

With the development of economy and information technology, a large amount of invoice information has been produced, the recognition of invoice information is urgent to realize its intelligent recognition. Most invoice issuing units basically adopt traditional manual identification methods for the processing of invoices. As the number of invoices increases, problems such as slow efficiency in identifying invoice information, error-prone, and difficulty in ensuring frequently appear (Zhi et al., 2021).

Giving the challenges and limitations of manual processing, this research aims to explore and evaluate Computer Vision (CV) approaches, such as, algorithms, methods, and techniques, with a strong focus on Artificial Intelligence (AI), to develop a robust and efficient system for automated information and extraction of key invoice information. The primary goal is to conduct a Systematic Literature Review (SLR) and evaluate the CV approaches, used in selected experiments.

Document comprehension is the automated process of reading, interpreting, and extracting information from written text and images contained within pages of documents (Subramani et al., 2020). For forms and structured documents with simple named entities (e.g., names, dates, prices), the existing extraction methods can already achieve a high accuracy. For unstructured documents and challenging content (e.g., addresses, tables, details), the automatic extraction is not yet good enough and still requires human assistance and validation (Riba et al., 2019).

Although there are already Layout Analysis (LA), Optical Character Recognition (OCR), and Information Extraction (IE) support technologies and even complete systems in the area of invoice recognition such as anyOCR (Mohsin Reza et al., 2018) propose, CloudScan (Palm et al., 2017), CUTIE (X. Zhao et al., 2019) and ORCMiner (Ha & Horák, 2022), described in the next chapter, fully automated systems that can recognize all the diverse information contained in invoices remain elusive.

In addition to the introduction and conclusion chapters, this study comprises four chapters: the theoretical background chapter, providing a comprehensive understanding of the fundamental concepts and terminologies related to invoice recognition; the

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

methodology chapter, outlines the systematic approach employed in this research, in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) guidelines. It details the preliminary research, objectives, Research Questions (RQ), the inclusion and exclusion criteria, search strategy, and the selection process for identifying and evaluating relevant reports.

The results chapter contains the findings of the systematic review, categorizes the various CV approaches used in invoice recognition and offering insights into the state of art in field. Additionally, this chapter highlights the key datasets used and the experimental details provided by the reports. The discussion chapter serves as the platform for in-depth analysis, interpretation, and contextualization of the research findings, addressing critical questions related to the invoice recognition process, identifying best practices, and providing a roadmap for future developments in the field of invoice recognition.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1 THEORICAL BACKGROUND

This chapter presents the theoretical concepts underlying the automatic invoice recognition process, namely the concept of an invoice (and some of its key elements), OCR, LA, IE, NER, and NLP. Some existing systems are also presented.

1.1 Invoice

An invoice is a documented record of a business transaction between a purchaser and a vendor, featuring a timestamp and a detailed breakdown of the items or services involved. Varieties of invoices encompass physical receipts, sales bills, debit notes, sales invoices, and digital online records (Adam Hayes, 2023).

1.1.1 Elements in invoices

This subsection, shows the specific elements that are typically found in invoices, including:

- Logo and branding elements.
- Header information.
- Costumer and Supplier information.
- Tables with described goods and services.
- Legal and Compliance information.
- Pricing information.

Each element of the invoice may require a different process to extract the key information.

1.1.2 Key information from invoices

In this subsection, are provided some key information from invoices, namely:

- Issue date.
- Invoice number.
- Supplier's name/domain.
- Supplier's Tax Identification Number (TIN).
- Costumer's TIN.
- Denomination and quantity of goods/services.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Transaction value.
- Value Added Tax (VAT).
- VAT amount.
- Reasons for not applying VAT.
- Company address.
- Costumer address.

Is important to extract this key information accurately from invoices, since there is information that can be easily misinterpreted, such as switching the company's address with the costumer's address.

1.2 Document comprehension systems

An effective and complete document comprehension system usually integrates Deep Learning (DL) based models, with multiple Deep Neural Networks (DNN) architectures for reading and understanding document content and should combine the fields of CV, Natural Language Process (NLP) and Named-Entity Recognition (NER) for a unified solution. A complete end-to-end system encompasses LA, OCR, and IE (Subramani et al., 2020).

1.2.1 Layout Analysis

Layout structure (or page segmentation) analysis refers to the process of segmenting a document page into groups of lines of text, blocks of text, and/or graphic elements (Ha & Horák, 2022). Instance segmentation methods provide labels by pixels to categorize regions of interest, such as text, pictures, images, and tables (Subramani et al., 2020).

1.2.2 Optical Character Recognition

An OCR model has purpose of locating and transcribing all the written text present in a document, having two main components, namely, text detection and text transcription. Generally, these two components are separate and use different models (Subramani et al., 2020).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1.2.2.1 Text Detection

Text detection requires separating the text region from the non-text region (Özgen et al., 2018). There are two types of text detection, Text Detection as object detection that provides information for semantic understanding of images, to determine where objects are in a given image (object location) and to which class each object belongs (object classification) (Z. Q. Zhao et al., 2019); and Text Detection as instance segmentation to solve the text density problem, which is the task of classifying each pixel of an image into specific and predefined categories (Subramani et al., 2020).

1.2.2.2 Word-level vs. Character level

Most systems try to directly detect words or even lines of words, some studies argue that character-level detection is an easier problem that text detection because characters are less ambiguous than lines of text or words (Subramani et al., 2020).

1.2.2.3 Text Transcription

Text transcription is the task of transcribing the text present in an image. The input, an image, is clipped, containing character, words, or sequence of words. A text transcription model processes this clipped image and produces a sequence of tokens. Predicting words instead of characters, makes typing mistakes less likely, such as substituting an ‘a’ for an ‘o’ (Subramani et al., 2020).

1.2.3 Information Extraction

IE aims to extract structures information from unstructured texts (Y. Lin et al., 2020). IE models use the OCR and/or LA output of documents to understand and identify relationships between information that is in the document.

1.2.3.1 Named-Entity Recognition

NLP focuses on the interactions between human language and computers. NLP is a way for computers to analyse, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

to perform tasks such as automatic summarization, translation, named entity recognition, and topic segmentation (Lopez & Kalita, 2017). NLP is used to perform IE.

NER, also known as entity identification or identity extraction, is a subtask of IE and is the task of extracting named entities, usually people or places, from unstructured text (Palm et al., 2017). In context where NLP is used for IE NER is a subtask of NLP.

Invoice content analysis relies heavily on being able to identify the relevant entities/typically person name, place name, organization name or product name). The NER task consists of two stages: entity identification and entity classification (Ha & Horák, 2022).

1.2.3.2 2D positional embedding

Sequence labelling increases the NER, incorporates 2D bounding boxes and combines them with text embeddings to create models that are simultaneously aware of the context and positional space of the information, via the x and y coordinate pair. The document is pre-processed to assign a line number to each token, which is assigned a sequential position. However, relying solely on the line number or the x and y coordinates of the bounding box can be misleading when the document is scanned on an uneven surface, leading to curved text (Subramani et al., 2020).

1.2.3.3 Image Embedding

IE for documents can also be framed as an CV challenge, where the objective of the template is to delimit bounding boxes in the areas of interest, to preserve the 2D layout of the document. While it is possible to learn strictly from the document image, image embedding simplifies the task of models to understand 2D textual relationships (Subramani et al., 2020).

1.2.3.4 Graph

Unstructured text in documents can also be represented as a graph network, where the nodes of graph represent different textual segments (Subramani et al., 2020).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1.2.3.5 Tables

Tabular data extraction continues to be a challenging aspect due to the variety of formats and similarity with other objects present in documents, for examples, charts, lists or flowcharts (Schreiber et al., 2017).

Tables are essential elements of documents for presenting structural information. Tabular data extraction consists of three main categories: tables detection, table recognition and comprehension. Table detection determines the table boundaries in an image. Table recognition looks for the internal structure of the table such as rows, columns, and table cells. Table comprehension refers to extracting the semantic content of the table (Rashid et al., 2017).

1.3 Invoice recognition automation systems

Below are presented some of the systems like the system that is proposed to be built, namely, anyOCR, CloudScan, CUTIE and ORCMiner:

1.3.1 anyOCR

Some elements in invoices, such as tables, header, footer, blank spaces, logo, among others, are typically not present on pages of books and magazines, which means that a standard LA system is inefficient on invoices. Mohsin Reza et al. (2018) propose a specific LA system for invoices, known as anyOCR. This system employs a line removal method to eliminate line graphics within tables, using binarization method, and combine lines of text so that the table information is kept intact line per line, text is recognized for each line in reading order and saved in hOCR format. In a performance comparison with standard anyOCR system and the commercial ABBYY system, anyOCR achieves an accuracy of 83.34%, surpassing 53.95% and 76.00%, respectively.

1.3.2 CloudScan

Palm et al. (2017) introduced CloudScan, a cloud-based invoice analysis system that requires no prior knowledge, meaning there is no need to configure to use it. Most systems require a minimum number of sample invoices, which limits the accuracy of unseen invoices. CloudScan does not rely on invoice layout templates, as it learns a single global

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

invoice template that naturally generalizes to unseen invoice layout, thereby extracting structured information from unstructured invoices. The model is trained with data automatically extracted from mandatory feedback provided by the user (Palm et al., 2017).

Here's how CloudScan works:

1. Takes an invoice in Portable Document Format (PDF) format as input.
2. Processes the text into n-grams for each line.
3. Selects n-grams relevant to key information.
4. Generates a Universal Business Language (UBL) invoice.
5. The UBL invoice and the original PDF invoice are presented to the user in a Graphical User Interface (GUI).
6. The user can make correct to the UBL invoice, which are the used to improve the system.

The Oracle classifier trained on n-grams and labels, extracted from validated UBL invoices and corresponding PDF's. When n-grams don't match any field, they are labelled as 'undefined'. If an n-grams matches multiple fields, it's assigned to all matching fields (Palm et al., 2017).

CloudScan's performance was evaluated through experiments. In these experiments, 70% of the invoices were used for training, 10% for validation and 20% for testing. The system extracts 32 key pieces of information, and its performance is measured by comparing the generated and validated UBL.

The Recurrent Neural Network (RNN) model, specifically Long Short-Term Memory (LSTM), and Baseline model (Logistic Regression) achieved F1-Score of 0.891 and 0.887, respectively, on seen invoice layouts. On unseen invoices, the RNN model outperforms Baseline model with an F1-Score of 0.840 compared to 0.788.

To improve the performance of the CloudScan system, it is necessary to improve the Oracle classifier and the discrepancies between UBL and PDF (Palm et al., 2017).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1.3.3 CUTIE

X. Zhao et al. (2019) introduce the Convolutional Universal Text Information Extraction (CUTIE) system, designed for IE. CUTIE employs a Graph Convolutional Network (GCN) DL model, which does not require pre-training or post-processing. Text graphs are formed with the grid mapping method. CUTIE allows to simultaneously query semantic and spatial information from the texts in the image of the scanned document and does not require a large dataset (X. Zhao et al., 2019).

To prepare input grid data, CUTIE processes scanned document image with OCR to obtain the texts and their absolute and relative position. The goal is to map the texts from the image to a destination grid while preserving the original spatial relationship of grid mapping texts. Grid data augmentation improves CUTIE's ability to handle documents with different layout (X. Zhao et al., 2019).

The experiment is evaluated on the SROIE dataset and on a self-build dataset with 3 types of scanned document images, namely, taxi receipts, meal receipts and hotel receipts, which contain 9 classes of key information, for each class, several tokens can be included.

The study compares two different networks architectures, CUTIE-A, a high-capacity CNN that merges multi-resolution features without losing high-resolution features, and CUTIE-B, an atrous convolution to expand the field of view and capturing multi-scale contexts with Atrous Spatial Pyramid Pooling (ASPP).

For the self-constructed dataset, both the softAP of CUTIE-A and CUTIE-B exceed their AP by a large margin. CUTIE-A achieves 90.8% AP and 97.2% softAP on taxi receipts, 77.7% AP and 91.4% softAP on meal receipts, and 69.5% AP and 87.8% softAP on hotel receipts. Compared to the CloudScan system, CUTIE-A and CUTIE-B outperform all test cases. Furthermore, compared to the Bidirectional Encoder Representations from Transformers (BERT) system for NER, CUTIE-A improves AP by 2.7% on taxi receipts, but is less accurate on the remaining types of documents; CUTIE-B improves AP by 5.9% on taxi receipts, 1.4% on meal receipts and 2.9% on hotel receipts, thus outperforming all test cases, being less complex and not requiring a large dataset for pre-workout. Regarding semantic information capacity, the best performance is achieved by CUTIE-B

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

with 85.0% in AP and 92.9% in softAP. The CUTIE model achieved good performance even with limited semantic information capacity (X. Zhao et al., 2019).

To evaluate the impact of the training dataset size, CUTIE was trained with different percentages of the dataset. CUTIE-B achieves the highest AP of 85% with 66% of the training dataset and the highest softAP of 91.5% with 75% of the training dataset. CUTIE-B is already able to reach 79.6% AP and 88.7% softAP with only 21% of the training dataset, which proves the efficiency of the proposed method (X. Zhao et al., 2019).

In another experiment, which includes a dataset of 613 images, two CUTIE-B models are trained to predict whether a text token is in a table region or not, and which table column is belongs to, one of the models locates the table region in the image, having located 86% of the tokens, and the other model identifies the specific column to which the token belongs, obtaining a result of 94.8% (X. Zhao et al., 2019).

The performance gain is achieved by three factors: the spatial relationship between the texts, the semantic information and the grid mapping mechanism (X. Zhao et al., 2019).

1.3.4 OCRMiner

Ha & Horák, (2022) present the OCRMiner system, for IE of scanned document images, capable of locating key information around keywords, which allows the system to adapt to layout varieties. In OCRMiner, the invoice image is processed through OCR, and the document layout is constructed based on key information and character placement. Through NLP techniques, annotations ate applied to the Extensible Markup Language (XML) file, which provide semantic and positional information, to create a context. After deciding the layout of the document, each block is categorized and given a position, including the absolute position on the page and the relative position of the block relative to other blocks (Ha & Horák, 2022).

In this study, XML files contain positional information of 7.602 blocks, 18.672 lines and 9.096 automated annotations of 64 annotation types. In evaluating the system's performance, the extracted key information is classified into 3 classes of results: full match, partial match and mismatch. The OCRMiner system was evaluated with two datasets, invoices in English and invoices in Czech. Using an open-source OCR, with a

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

dataset composed of 20 invoices, the system can detect 90.5% of elements of the English dataset and 88.2% of Czech dataset, with 93%-96% of these elements being correctly identified. OCRMiner is also able to detect whether an image is the first page of an invoice with an accuracy of 98% (Ha & Horák, 2022).

Both datasets have high match rates on page number, invoice date, and due date. For the English dataset, the results are better for the invoice number, the VAT number, the Identification Bank Account Number (IBAN), the total amount. In contrast, the Czech dataset performs better with order number, account number, payment date. There is a poor match for the payment date in the English dataset due to OCR errors. Title Identification (ID) and page number achieve 96% accuracy. Among the classifiers, Naïve Bayes achieves the highest recall of 99%, while Logistic Regression achieves the best balance between accuracy and recall with an F1-Score of 98% (Ha & Horák, 2022).

To compare the performance with another system, the OCRMiner dataset was evaluated with the InvoiceNet system, which uses DNN models. OCRMiner generally performs better than InvoiceNet on all key information selected for the experiment. InvoiceNet can more precisely locate the position of the addresses in the English dataset, but it fails to extract the complete information (Ha & Horák, 2022).

OCRMiner system cannot accurately distinguish the address of the supplier and the costumer, which requires manual intervention, in the blocks related to the supplier, there are 27% errors in the English dataset and 35% errors int the Czech dataset. The Czech dataset causes 17% of layout errors compared to only 3% for the English dataset. Error analysis of partial matches revels that OCR errors cause 64% of errors for the English dataset and 34% for the Czech dataset. The incomplete address is found in 46% of the English dataset and 16% in the Czech dataset (Ha & Horák, 2022).

To improve the performance of OCRMiner, a model is needed that differentiates the information related to the supplier and the costumer, improving the performance of the LA. Identification of remote information such as supplier and buyer address can further increase the accuracy of extracted information (Ha & Horák, 2022).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2 METHODOLOGY

In this section, we present the methodology following the PRISMA framework, that ensures a systematic and transparent approach to identify, select, and evaluate studies related to invoice recognition, allowing to produce a comprehensive and reliable systematic review (Novais, 2023; Page et al., 2021).

2.1 Preliminary research

Before applying the SLR method, we conducted preliminary research to gain an understanding of the study's domain. We produced a Word Cloud in Python, as shown in Figure 1, where the words document, model, text, image, word, method, and neural network stand out, which point to the use of image and text processing models and methods. The neural network bigram identified at the end of the figure stands out. From the results of this first cloud, as will be seen below, the research focused on the use of the words Text Detection, Deep Learning, and Invoice (which was not very prominent in this cloud, but which is the subject of the study).

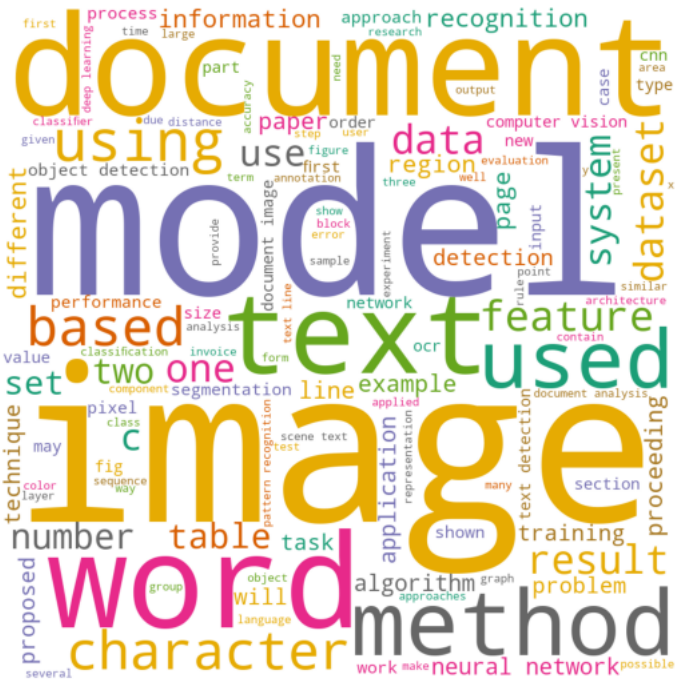

Figure 1. Word Cloud for SLR

Source: Own work, made with Python

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

2.2 Objectives and Research Questions

In the first survey conducted to gain an understanding of concepts related to automatic invoice recognition systems, it became evident that the invoice recognition process comprises several stages and tasks, with CV being the primary technology employed throughout these stages and tasks.

Therefore, the objective of this SLR study is to identify and categorize the various CV approaches, including algorithms, techniques, and methods, and evaluate their effectiveness and performance based on empirical studies.

Our Research Questions (RQ) are as follows:

- RQ1: What are the essential stages involved in conducting an invoice recognition process?
    - RQ1.1: Why is it important to preprocess images?
    - RQ1.2: How does LA improve invoice recognition?
- RQ2: Which algorithms are commonly employed to execute distinct stages, as LA, OCR, and IE?
- RQ3: What software and hardware configurations are commonly used in these CV approaches?
- RQ4: Which programming languages are frequently utilized in the development of invoice recognition process?
- RQ5: Which datasets are typically employed for research purposes within each stage of invoice recognition process?

2.3 Inclusion and Exclusion Criteria

In this section, we outline the criteria used to determine the inclusion and exclusion of articles in the review.

Inclusion Criteria:

- Studies published after the year 2017.
- Studies that specifically report on experimental research.
- Studies with a comprehensive description of model architecture and parameters.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Studies that, due to data privacy concerns, use datasets other than invoices.
- Studies that have comparative analysis of different CV approaches.
- Studies that address data preprocessing techniques.
- Studies that discuss the computer infrastructure utilized.

Exclusion Criteria:

- Duplicate publications.
- Non-open access publications.
- Studies with inadequate experimental details.
- Studies that predominantly focus on theoretical frameworks or historical developments.
- Studies that do not primarily concentrate on AI techniques.
- Non-English language publications.

2.4 Search Strategy

The search strategy employed for the SLR, leverages B-on (Online Knowledge Library) database as primary source for retrieving relevant academic articles and research reports. This section provides an overview of the key search terms selected for the advanced search and specifies the publication year of the reports.

2.4.1 Information sources

In the systematic review, the search strategy was conducted using B-on databases, that provides access to a wide range of academic and research databases, which encompass a variety of scientific disciplines. The B-on database, provides access to a comprehensive selection of scientific resources and research repositories, enabling a thorough exploration of relevant literature (Novais, 2023).

2.4.2 Search terms

In this SLR, three primary keywords were chosen:

- **Invoice**: Given that the study focuses exclusively on invoice documents.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- **Text Detection**: As the core and most intricate aspect of the study involves identifying characters within the documents.
- **Deep Learning**: Originally, the intention was to utilize the term 'Neural Networks'. However, it was noted that many algorithms incorporate the phrase 'Neural Networks' in their names, which could potentially affect the accuracy of the research. Therefore, 'Deep Learning' was chosen instead.

The search expression that combines these selected keywords and the year criteria is as follows:

ALL(INVOICE) AND ALL(TEXT DETECTION) AND ALL(DEEP LEARNING) AND (YEAR > 2017) AND (DATE <= 2023-01-09)

2.5 Selection process

For the SLR, the B-on's database was employed, functioning as an aggregator of academic resources similar to platforms such as Scopus, IEEE Xplore, Web of Science, and Springer, to select relevant reports.

The initial search was made on January 9th, 2023. Since B-on does not allow specifying the publication year in advanced searches, the initial query was executed without specifying the year, resulting in 25.795 records, as depicted in Figure 2.

Figure 2. Advanced search using B-on

Subsequently, the search was refined by utilizing que year filter available in B-on's filtering parameters, as shown in Figure 3. This narrowed it down to 10.768 records matching the search string presented in the previous section, meaning that the reports are from January 1st, 2017, to January 9th, 2023.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Figure 3. Advanced Search in B-on, for year specification

The source of content selected was Academic Journals, which generated 342 results. Further refinement reduced the number of articles to 182 by selecting the subjects that appeared most frequently in the Word Cloud, such as deep learning, machine learning, artificial intelligence, convolutional neural networks, artificial neural networks, feature extraction, natural language processing, classification, data mining, big data, support vector machines, text recognition, image processing, supervised learning, computing and processing, image segmentation, data analysis, computer science, visualization, literature review, decision making, object recognition, image analysis, image recognition, object detection, databases, classification algorithms, text mining, natural languages, detectors and data augmentation.

After the filtering process, the following search string was obtained:

ALL(INVOICE) AND ALL(TEXT DETECTION) AND ALL(DEEP LEARNING) AND (YEAR > 2017) AND (DATE <= 2023-01-09) AND DOCUMENT TYPE(ACADEMIC JOURNALS) AND SUBJECT(DEEP LEARNING, MACHINE LEARNING, ARTIFICIAL INTELLIGENCE, CONVOLUTIONAL NEURAL NETWORK, ARTIFICIAL NEUTAL NETWORK, FEATURE EXTRACTION, NATURAL LANGUAGE PROCESSING, CLASSIFICATION, DATA MINING, BIG DATA, SUPPORT VECTOR MACHINES, TEXT RECOGNITION, IMAGE PROCESSING, SUPERVISED LEARNING, COMPUTING AND PROCESSING, IMAGE SEGMENTATION, DATA ANALYSIS, COMPUTER SCIENCE, VISUALIZATION, LITERATURE REVIEW, DECISION MAKING,OBJECT RECOGNITION, IMAGE ANALYSIS, IMAGE RECOGNITION, OBJECT DETECTION, DATABASES, CLASSIFICATION ALGORITHMS, TEXT MINING, NATURAL LANGUAGES, DETECTORS, DATA AUGMENTATION)

After reviewing the title and the abstract of each of the 182 reports, only 61 were selected for analysis because they were found to be focused on the research topic. This selection was made even though they contained the search keywords and met the previously described filters. Reports related to medical themes were excluded.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

However, the analysis was further limited to only 20 reports. This was because some reports could not be located, while others were surveys or primarily focused on statistical analysis rather that experimental research related to CV approaches.

Out of the 20 reports analysed, there were a few that were not directly relevant to the topic under investigation. However, since they had some tangential connection to the subject matter, these reports were included in APPENDIX 2.

A visual representation of the entire review process can be found in Figure 4.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

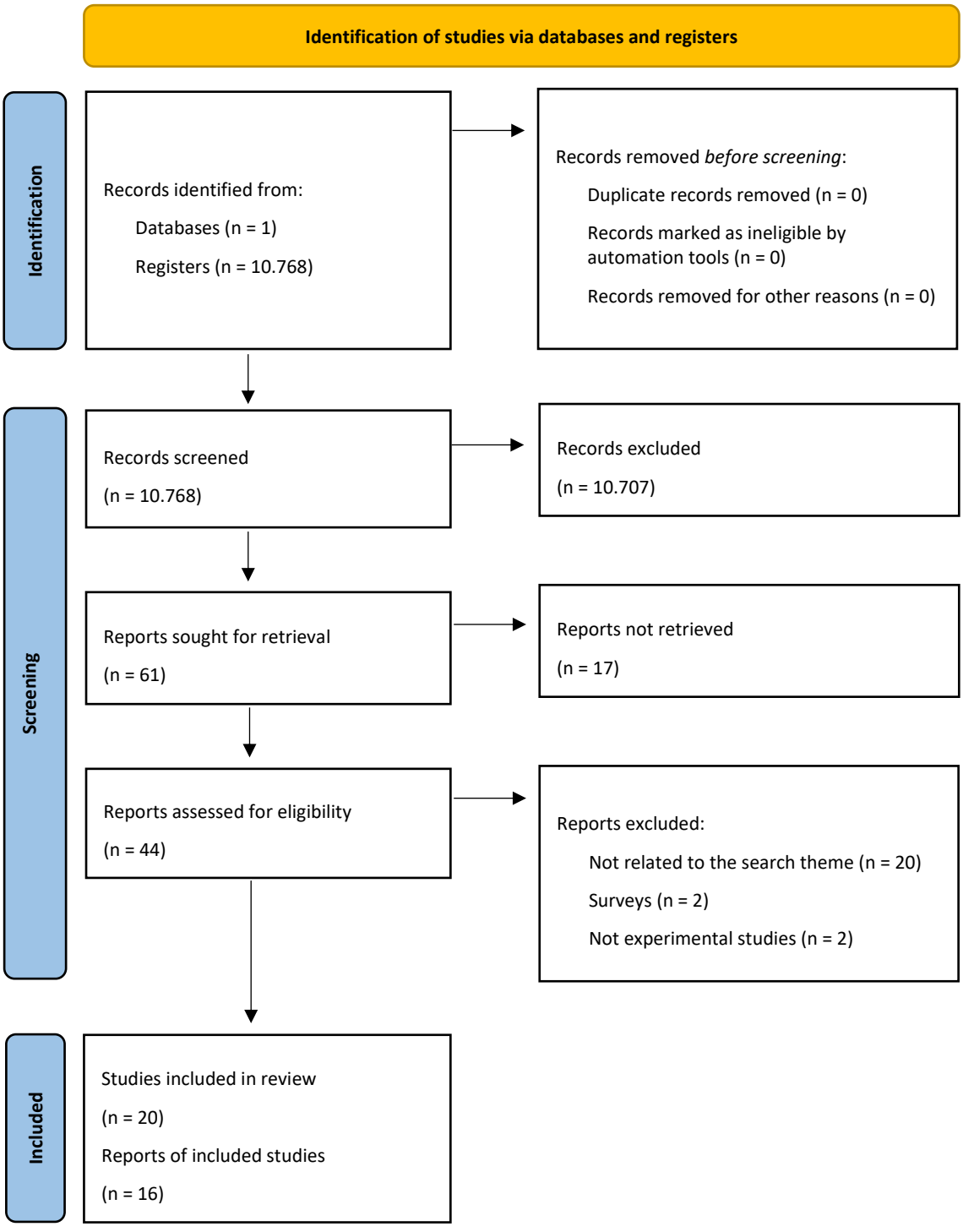

Figure 4. PRISMA Flow Diagram for the SLR

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

3 RESULTS

The chosen reports are experiments, and the objective is to extract the essential information related to CV approaches. These studies were selected for their relevance to the research focus and their potential contributions to the understanding of CV techniques.

3.1 Description of selected articles

The selected studies are briefly presented below, each offering unique insights into the realm of CV.

3.1.1 A CCD based machine vision system for real-time text detection

S. Zhao et al. (2020) proposed a Charge-Coupled Device (CCD) based machine vision system for real-time text detection in invoice images. This system applies optimizations from various aspects, including the optical system, the hardware architecture, and DL algorithm, to enhance the speed performance of the machine vision system. The Connectionist Text Proposal Network (CTPN) algorithm was utilized in the experiment. The CTPN pipeline consist of four stages: Convolutional Neural Networks (CNN), LSTM, Proposal, and Detection.

- Algorithms: CTPN, based on F-RCNN, was tested with four different architectures: VGG16, InceptionV3, ResNet50 and MobileNet.
- Software/Hardware Configuration: CPU: Intel i7; GPU: NVIDIA GTX1080.
- Programming Language: Not available; Code: not available.
- Dataset(s): Private dataset of invoice documents (not available).

CTPN demonstrates excellent detection performance for various general scene images and is capable of detecting text lines in invoice images. Superior performance is achieved by combining prior knowledge of the application scenario with appropriate preprocessing algorithms. The time consumption of the CTPN was tested in various hardware environment. In the Central Process Unit (CPU) environment, the time consumption averages around 8 seconds per image, which can be reduced to 4 seconds with Single Instruction Multiple Data (SIMD) utilization. In the Graphic Processing Unit (GPU)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

environment, the time consumption decreases significantly to approximately 0.1 to 0.4 seconds per image.

To evaluate the impact of the network model on algorithm speed performance, the MobileNet network architecture demonstrated a significant reduction in time consumption. Experimental data confirms that optimization methods such as SIMD acceleration, GPU acceleration and improvement in the optical system, effectively enhance the operational speed of the machine vision system. This optimization ensures that the system meets the real-time text detection requirements in industrial scenarios.

3.1.2 A method for identifying the key information of electronic invoicing in complex scenes

Zhi et al. (2021) present a novel Knowledge-Based Key Information Detection and Recognition (KBKIDR) method for electronic invoices in complex scenes. During the training stage, complex scenes are simulated using data augmentation techniques such as addition of random noise (Gaussian noise and salt-and pepper), colour jitter, horizontal lines, and random rotation to enhance text recognition accuracy. In the modelling stage, the proposed method leverages prior knowledge to extract key information, thereby improving recognition processing efficiency.

- Algorithms: CRNN, CNN+GRU[1], CONV+FC[2], DN[3]+GRU, CRNN, CNN+GRU, CONV+FC and DN+GRU.
- Software/Hardware Configuration: Windows 10 (64-bit); CPU: Intel(R) Core (TM) i7-10510U (1.8GHz); RAM: 16GB.
- Programming language: Python 3.6.7; Code: Not available.
- Dataset(s): Private dataset (not available).

The experiment results demonstrate that the proposed method exhibits reliable recognition accuracy for the key information in electronic invoices in complex scenes and significantly improves the recognition efficiency. In a comparative experiment for image

[1] Gate Recurrent Unit
[2] Fully Connection
[3] Dense Net

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

enhancement, CRNN has the higher accuracy at 99.03%. In the image slicing experiment, CRNN had the lowest average time consumption, recording 3.682 seconds.

3.1.3 A multi-pronged accurate approach to optical character recognition, using nearest neighbourhood and neural-network-based principles

G KISHOR KUMAR et al. (2021) proposed a Multi-Layer Perceptron (MLP) neural network architecture that includes an input layer, hidden layers, and an output layer to create an effective OCR method. The architecture constructs a model that learns data representation from input data and employs these representations for classifying unknown data. The performance of MLP, specialized in OCR, is compared with existing nearest neighbour methods such as Condensed Nearest Neighbour (Cond NN[4]), Modified Condensed Nearest Neighbour (MCNN) and Other Class Nearest Neighbour (OCNN).

- Algorithms: Cond NN, OCNN, MCNN and MLP.
- Software/Hardware Configuration: Keras library; CPU: i3 core; RAM: 4GB.
- Programming Language: Python; Code: Not available.
- Dataset(s): OCR (Susheela Devi & Murty, 2002) and Pendigits (Dua & Graff, 2019).

MLP demonstrates superior performance compared to Cond NN, MCNN and OCNN, achieving an accuracy of 96.01% on the OCR dataset and 99.21% on the Pendigits dataset.

3.1.4 All-content text recognition method for financial ticket images

Zhang et al. (2022) have developed an all-encompassing method for detecting and recognizing text information in financial ticket images based on DL. This method effectively mitigates common image noise and enables batch extraction of financial data from ticket images. To address the challenge of recognizing mixed character in a multi-character context, they propose a Financial Ticket Character Recognition Framework (FTCRF), which comprises three key components: Pixel Aggregation Network (PAN) as

[4] The abbreviation Cond NN is used because it is not possible to use the abbreviation CNN already in use for Convolutional Neural Network.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

a text region detection model; F-RCNN for character segmentation; and Double Focal Loss - Convolutional Neural Network (DFL-CNN) for character recognition. The FTCRF significantly improves the accuracy of recognizing mixed character and streamlines the detection and recognition of financial ticket information.

- Algorithms: For text detection: PAN-512, PAN-640, EAST (Efficient and Accurate Scene Text Detector) and CTPN; For character segmentation: F-RCNN, Connected Components, Projection and improved F-RCNN; For character recognition: DN-121, ResNeSt-101, EfficientNet-b7 and DFL-CNN.
- Software/Hardware Configuration: TensorFlow and PyTorch; CentOS Linux 7 Core; GPU: 32G Tesla V100 (2x); CPU: 2.20 GHz; Memory: 256 GB.
- Programming Language: Python; Code: Not available.
- Dataset(s): Private dataset of bank tickets (not available).

In text detection, the PAN-640 model stands out for its superior accuracy and speed. The proposed method is compared to the F-RCNN algorithm, Connected Components and Projections, outperforming them with a higher Recall rate and F1-Score, 0.99 and 0.96. When evaluating four depth model classifiers, DFL-CNN emerges as the top performer due to its attention to image details and sensitivity to different characters. The experimental results showcase the effectiveness of the PAN + FTCRF method, achieving a string recognition accuracy of 91.75%, and single-ticket processing time of 578.33 ms. The overall recognition accuracy for entire tickets reaches 87%, significantly enhancing the efficiency of financial accounting system.

3.1.5 Automatic Receipt Recognition System Based on Artificial Intelligence Technology

C. J. Lin et al. (2022) have introduced an Automatic Receipt Recognition System (ARRS). The system's workflow involves scanning receipts to create high-resolution images. Receipt characters are categorized into two groups based on their characteristics: printed and handwritten characters. Subsequently, different preprocessing techniques are applied to each category.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

For handwritten characters: Text positioning is achieved using template matching and fixed receipt features, character segmentation is accomplished through projection techniques and character recognition if performed using a CNN.

For printed characters, the proposed You Only Look Once version 4-small (YOLOv4-s) model is employed for precise text positioning and character recognition.

- Algorithms: CNN, YOLOv4-s and YOLOv4.
- Software/Hardware Configuration: Not available.
- Programming Language: Not specified; Code: Not available.
- Dataset(s): Not available.

Experimental results demonstrate that the CNN achieved an 80.93% recognition accuracy for handwritten characters, while the YOLOv4-s model had a 99.39% accuracy rate for printed characters. Notably, the recognition accuracy of the YOLOv4-s model surpassed the traditional YOLOv4 model by 20.57%. The proposed ARRS presents a substantial enhancement in tax declaration efficiency, cost reduction, and simplification of operational procedures.

3.1.6 Beyond document object detection: instance-level segmentation of complex layouts

Biswas et al. (2021) focus on going beyond object detection for understanding document layout. The authors add another segmentation module to state-of-the-art document object detection system that can generate segmentation masks for every individual object category of a document image. Their proposal introduces an end-to-end instance-level segmentation model inspired by the state-of-the-art instance segmentation models, Mask-RCNN and Mask Scoring Region-based Convolutional Neural Networks (MS-RCNN). The main purpose is to detect layout objects such as tables, figures, paragraphs, and titles.

- Algorithms: Mask-RCNN and MS-RCNN models, with ResNeXt-101 as a backbone, Feature Pyramid Network (FPN).
- Software/Hardware Configuration: PyTorch library; GPU: Nvidia Titan X.
- Programming Language: Python; Code: Not available.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Dataset(s): PubLayNet (Zhong et al., 2019) and HJDataset (Shen et al., 2020) + ImageNet (Studer et al., 2019) to train ResNeXt-101.

ResNeXt-101 was chosen as the backbone over ResNet-101 for its superior mean Average Precision (mAP). Performance analysis conducted on the PubLayNet dataset reveals that the model M-RCNN (+ FPN) attains an mAP of 0.904. For both document object detection and document instance segmentation tasks, the proposed method achieves the higher AP, 0.920 and 0.893, respectively. Results for the HJDataset also demonstrate the effectiveness of the proposed model, with an AP of 0.822 for document object detection and 0.820 for instance segmentation.

3.1.7 Denoising Letter Images from Scanned Invoices Using Stacked Autoencoders

In this paper, Alshathri et al. (2022) focus on the preprocessing images to enhance the quality of scanned invoice images before applying OCR. The letter data extracted from invoice images undergo denoising using a modified autoencoder-based DL method. They employ a Stacked Denoising Autoencoder (SDAE) with two hidden layers in in both the encoder and decoder network. An undercomplete autoencoder is designed with non-linear encoder and decoder function to capture the most salient features from the training samples. This autoencoder is regularized for denoising application using a combined loss function that considers both mean square error and binary cross entropy. Performance is analysed in terms of Signal to Noise Ratio (SNR), Peak Signal to Noise Ratio (PSNR), Structural Similarity Index (SSIM) and Universal Image Quality Index (UQI) and the results are compared with other filtering techniques like Nonlocal Means filter, Anisotropic diffusion filter, Gaussian filters, and Mean filters.

- Algorithms: SDAE, SNR, PSNR, SSIM and UQI.
- Software/Hardware Configuration: PyTorch library; GPU: NVIDIA Tesla k20M.
- Programming Language: Python; Code: Not available.
- Dataset(s): 59,119 letter images, with English alphabets and numbers from many scanned invoices images (not available).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

The denoising performance of proposed SDAE method is compared with existing SDAE methods employing a single loss function in terms of SNR and PSNR values. The results show the superior performance of the proposed SDAE method.

3.1.8 End to End Invoice Processing Application Based on Key Fields Extraction

Arslan, (2022) presents an automated invoice processing system, capable of handling various invoice file types. Companies can easily submit invoices through the system's web interface or email, and all submitted invoices are queued and processed sequentially. Depending on the format of the invoice, different extraction methods are employed. For text-based invoices, invoice information is extracted using template matching. In cases where invoices are in image format, both text and table areas are detected and extracted. The system utilizes image processing techniques, including morphological operations, as well as the You Only Look Once version 5 (YOLOv5) and M-RCNN algorithms for table detection.

- Algorithms: M-RCNN and YOLOv5.
- Software/Hardware Configuration: OpenCV, Tesseract 4.0 with Pytesseract; Docker; CPU: Intel Core i7-11800H CPU; GPU: NVIDIA GeForce RTX 3050 Ti; RAM: Memory: 16 GB.
- Programming Language: React, Java and Python; Code: Not available.
- Dataset(s): Turkish (TUR) and English (ENG) invoice documents (not available).

Experimental results for table detection reveal an mAP of 95.14% with image processing techniques, 98.07% with YOLOv5 and 98.19% with M-RCNN. YOLOv5 was chosen due to its faster processing and lower system requirements. Pre-trained models from Tesseract, including TUR and ENG, were employed for Turkish and English documents. However, since TUR model's accuracy was insufficient, a new model, named TTUR, was trained. The experimental results demonstrate that the TTUR model achieved higher accuracy at 90.35% compared to TUR, 83.08%, and ENG, 84.70%, models provided by Tesseract.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

3.1.9 Form location and extraction based on deep learning.

Z. Zhang et al. (2020) present a form location and extraction method based on an improved F-RCNN to locate and extract information. They employ two classification models to enhance detection accuracy: model A is used to detect the smallest rectangular area containing the main information, while model B is designed for recognizing usernames, phone numbers and addresses.

- Algorithms: F-RCNN with RPN, F-RCNN and You Only Look Once version 3 (YOLOv3).
- Software/Hardware Configuration: MATLAB 2018b; UBUNTU 16.04.5.; CPU: Intel Xeon E5-1620 v4 @ 3.50GHz; GPU: NVIDIA TITAN Xp, 12GB GDDR5; RAM: 64GB DDR4.
- Programming Language: Matlab; Code: Not available.
- Dataset(s): Private dataset (not available).

Experimental results demonstrate that the proposed approach, utilizing cascade models, achieves high accuracy in processing waybill images with complex background. The system's performance is significantly enhanced by the suggested anchor sizes and the positive sample selecting method, resulting in an 82.44% accuracy. Object detection attains the highest mAP, with phone number displaying higher accuracy compared to name or address due to their fixed length.

3.1.10 Fusion of visual representations for multimodal information extraction from unstructured transactional documents

Oral & Eryiğit, (2022) investigate the impact of employing different visual representations and their fusion on information extraction from unstructured transactional documents, particularly focusing on complex relation extraction from money transfer order documents. They introduce and experiments with five distinct visual representation approaches: word bounding box, grid embedding, grid convolutional neural network, layout embedding, and LGCN. Additionally, they explore various fusion strategies, including three basic vector operations, weighted fusion, and attention-based fusion.

The information extraction architecture used in this study consists of the following stages:

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

1. OCR: Extract texts within documents.
2. Document classification: Document are classified based on their process types.
3. Information extraction: Tailored to the specific process types and structures.
4. Post-processing: Automatic and human validations of the extracted transaction information.

The information extraction stage comprises two sequential parts: a NER stage and complex relation extraction stage.

- Algorithms: G-CNN and LGCN; Hyperparameters of Bidirectional Long Short-Term Memory (BiLSTM) and MLP.
- Software/Hardware Configuration: Not available.
- Programming Language: Not available; Code: Not available.
- Dataset(s): Unstructured money transfer order documents (not available).

Weighted fusion and attention-based fusion strategies yield relative error reductions of up to 30% eliminating the need to select appropriate visual approaches for task at hand. Attention-based fusion model is found to be the most successful approach on multi-transaction document understanding.

3.1.11 Invoice Detection and Recognition System Based on Deep Learning

Yao et al. (2022) designed and implemented an invoice information recognition system based on DL. The system addresses challenges posed by low image contrast, lack of image due to poor lighting, or noise effects by applying image preprocessing techniques such as image greying and normalization. Subsequently, a target detection and invoice recognition method is devised, combining YOLOv3 and CRNN. This approach results in an end-to-end invoice information recognition model used to develop a DL-based invoice detection and recognition system.

- Algorithms: YOLOv3 + CRNN.
- Software/Hardware Configuration: OpenCV 2.4.10 for image processing, LibXL for data logging to Excel and MuPDF for PDF parser; Windows 7, 64 bits; CPU: clocked at 2.0 MHz; RAM 4GB.
- Programming Language: C++; Code: Not available.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Dataset(s): Private dataset (not available).

Experiments resus validate the system's attributes, including high recognition accuracy and efficiency. The system demonstrates precise identification of invoice content information, contributing to reductions in resource utilization and manpower.

3.1.12 Recurrent Convolutional Neural Network MSER-Based Approach for Payable Document Processing

Aladhadh et al. (2021) introduce an end-to-end OCR system designed to streamline the processing of payable documents, such as cheques and cash disbursement. This system combines text localization and recognition within a single unit to automate the entire process.

For text localization, the Maximally Stable Extremal Region (MSER) is employed, extracting a word or digit segments from invoices. These segments are then passed to a DL model, integrating CNN and LSTM. The CNN serves for feature extraction, with the extraction features subsequently fed to the LSTM. This model unifies feature extraction, sequence modelling, and transcription, handling sequences of varying lengths independently of characters segmentation or horizontal scale normalization. Notably, it supports both lexicon-free and lexicon-based text recognition. Additionally, the model is optimized to be compact, suitable for practical applications.

MSER Pipeline consists of two phases:

1. Image detection: Identifying textual geometries in a scanned document.
2. Transcription: Converting detected textual geometries into machine-readable text.

- Algorithms: MSER, RCNN (CNN+LSTM), AdaDelta optimizer technique for Stochastic Gradient Descent (SGD).
- Software/Hardware Configuration: CPU: 20x; GPU: GeForce GTX 1080 Ti; RAM 12 GB.
- Programming Language: Not available; Code: Not available.
- Dataset(s): Private dataset (not available).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Results from the experiment evaluation showcase the effectiveness of the model, achieving an accuracy of 95% without lexicon, and 99% with lexicon. The model exhibits versability and can be applied to various similar recognition scenarios.

3.1.13 Research on fast text recognition method for financial ticket image

H. Zhang, Dong, Zheng, & Feng, (2022) analyse the distinct features found in financial tickets, categorizing them into three distinct groups. For each category, they propose specific recognition patterns capable of meeting the diverse requirements for financial ticket recognition.

In this study, they introduce a straightforward, yet effective network known as the Financial Ticket Faster Detection network (FTFDNet) which is based on F-RCNN. To enhance text recognition accuracy, the loss function, RPN, and Non-Maximum Suppression (NMS) are customized to prioritize text detection, capitalizing on the unique characteristics of financial ticker text.

- Algorithms: COCO2017 pre-trained model, FTFDNet based F-RCNN; ResNetSt101 as backbone, Inception-RPN for area detection; Improved RPN, and NMS.
- Software/Hardware Configuration: MMDetection and Detectron2 for object detection; GPU: Tesla P40 24GB.
- Programming Language: Not available; Code: Not available.
- Dataset(s): Private dataset (not available).

Experimental results validate the effectiveness of these enhancements. When using F-RCNN as the network framework, employing the ResNet variant, ResNetSt101, yields improved results. Using Inception-RPN, not only enhances recognition accuracy but also accelerates the recognition speed by 50%, surpassing the outcomes of related methods with a recognition accuracy of 97.4%.

3.1.14 Table Detection in Invoice Documents by Graph Neural Networks

Riba et al. (2019) propose a graph-based approach for detecting tables in document images. Instead of using the raw content, such as recognized text, this approach utilizes

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

location, context, and content type to focus solely on structure perception. The proposed framework makes use of Graph Neural Networks (GNNs) to capture local repetitive structural patterns in invoice documents. Node classification and Edge classification are applied to further refine the detection process.

- Algorithms: GNNs
- Software/Hardware Configuration: ABBYY FineReader.
- Programming Language: Python; Code: https://github.com/dhavalpotdar/Graph-Convolution-on-Structured-Documents/blob/master.
- Dataset(s): RVL-CDIP (Harley et al., 2015) and CON-ANONYM (private dataset).

One notable advantage of this approach is its ability to handle anonymized data, which is a significant concern for companies dealing with sensitive content like invoices. This is achieved without relying on textual content and has yielded promising results. For both datasets, the optimal threshold is found to be 0.1, resulting in an F1-Score of 73.7 for CON-ANONYM, and 30.8 for the RVL-CDIP dataset.

3.1.15 Table information extraction and analysis: A robust geometric approach based on GatedGCN

Liu et al. (2022) propose a novel model to extract key information from documents and reconstruct table information from structured images, using GatedGCNs, this model considers three kinds of features for the semantic entities, including the position of an entity, the box containing the entity and texts inside the box. The model also considers the relationship between semantic entities, which is a key factor to improve the classification accuracy.

- Algorithms: GatedGCNs.
- Software/Hardware Configuration: GPU: NVIDIA Tesla V100.
- Programming Language: Not available; Code: Not available.
- Dataset(s): Medical Invoice (not available), Train Tickets (not available), SciTSR (Chi et al., 2019), ICDAR2013 dataset (Gobel et al., 2013).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

By combining the extracted node and the edge features, this model demonstrates outstanding performance when applied to key field extraction and table reconstruction. Its overall results exhibit high precision, recall, F1-Score, and accuracy are all surpassing 0.84.

3.1.16 Table Localization and Segmentation using GAN and CNN

Tables within documents often pose challenges due to the proximity of rows or columns, and in some cases, overlapping columns. Traditional techniques relying on hand-engineered features struggle to generalize because they are not adaptable to varying layouts. Recently, DL approaches have shown promise in table localization and segmentation tasks. In this paper, Reza et al. (2019) present an innovative approach using a conditional Generative Adversarial Networks (cGAN) based architecture for table area localization and SegNet based encoder-decoder with skip connections architecture for table structure segmentation.

- Algorithms: cGAN; SegNet.
- Software/Hardware Configuration: PyTorch and OpenCV libraries.
- Programming Language: Python; Code: Not available.
- Dataset(s): ICDAR 2013 (Gobel et al., 2013).

Result reveal that the pix2pixHD architecture for table localization achieves remarkable performance with an accuracy of 98.29%. For table segmentation, the SegNet based Encoder-Decoder with skip connections achieves 96% accuracy in segmenting rows and columns.

3.2 Summary of results

Before presenting the summarized results, it's important to provide an overview of how these results were compiled, considering the research questions to be answered, which are analysed in the subsequent discussion chapter.

In the previous section, we conducted data coding to categorize and identify essential information from each of the selected reports. This information is crucial for future endeavours in automatic invoice recognition, aligning with goals of our study. The

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

primary focus during this research phase was on the algorithms used. Additionally, some authors employed specific architectural configuration, including layers and connections, to enhance model performance. Various techniques were also identified, such as image preprocessing and the application of one or more algorithms.

Across the reviewed papers, a standard process emerged for extracting key information, particularly regarding Preprocessing, OCR, LA, and IE, corresponding to our first RQ. However, it's important to note that not all these stages are always necessary, and their applicability varies based on specific use cases. Furthermore, some of the identified stages also have tasks and sub-tasks. The following structure outlines the stages, tasks, and sub-tasks. Stages identified in the SLR that should be considered in a future invoice recognition system:

- Preprocessing
- LA:
    - General
    - Instance Segmentation
- OCR:
    - General
    - Text Detection
        - as object detection
    - Text Localization
    - Text Recognition
    - Character Level
        - Character Segmentation
        - Character Recognition
    - End-to-end model
- IE:
    - General
    - Unstructured Documents
    - Tables
        - General

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Table Detection
- Table Localization
- Table Segmentation
- Table Recognition

This categorization of stages, tasks and subtasks will be used to create some of the tables that present the results in the following sections.

The summary of the results presented in the following sections will help answer the RQ, which are answered in the discussion chapter.

3.2.1 Summary of Computer Vision approaches

After systematizing and identifying the different stages, tasks and associated sub-tasks presented above, we move on to identifying the algorithms, techniques and methods used, presenting in this section a summary of the algorithms found in the different tables.

3.2.1.1 Layout Analysis

For this stage of the process, the results presented in the Table 1were obtained.

Table 1. Algorithms for LA identified in SLR

Task	Sub-task	Algorithms	Results	References
Instance Segmentation		Model is inspired on M-RCNN and MS-RCNN models and adapted the ResNeXt-101 as a backbone, FPN.	Combination of M-RCNN and FPN achieved a higher mAP, of 0.904. The proposed method demonstrates an AP exceeding 0.820.	Biswas et al. (2021)

3.2.1.2 Optical Character Recognition

For this stage of the process, the results shown in the Table 2 were obtained.

Table 2. Algorithms for OCR identified in SLR

Task	Sub-task	Algorithms	Results	References
Text Detection	**As Object Detection**	CTPN and Architectures: VGG16, InceptionV3, ResNet50 and MobileNet.	MobileNet decreases time consumption and hardware acceleration and optical system improvement improve running speed of the system.	(S. Zhao et al., 2020)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

		PAN-512 and PAN-640, EAST and CTPN.	PAN-640 has advantages in accuracy and speed.	(H. Zhang et al., 2022)
		Proposed model is Inspired on F-RCNN and M-RCNN with adapted ResNetXt-101 backbone.	Proposed model shows higher AP than the original models F-RCNN and M-RCNN.	(Biswas et al., 2021)
		COCO2017 as pre-trained model, FTFDNet based F-RCNN and Improved RPN; ResNeSt101 as backbone and Inception-RPN for area detection.	97.4% of accuracy and the recognition speed increases by 50%.	(H. Zhang et al., 2022)
Text Localization		MSER	Robustness to noise and illumination and detects text chunks.	(Aladhadh et al., 2021)
Text Recognition		CRNN, CNN+GRU, CONV+FC, DN+GRU, CRNN, CNN+GRU, CONV+FC and DN+GRU.	With CRNN algorithm, accuracy reaches 99.03% with image enhancement and minimal time consumption when using image slicing.	(Zhi et al., 2021)
		RCNN (CNN+LSTM).	95% accuracy without lexicon and 99% with lexicon.	(Aladhadh et al., 2021)
Character level	**Character Segmentation**	Improved F-RCNN, Connected Components, Projection and improved F-RCNN.	Recall and F1-Score is higher in the improved F-RCNN model.	(H. Zhang et al., 2022)
	Character Recognition	DenseNet-121, ResNeSt-101, EfficientNet-b7 and DFL-CNN architectures.	DFL-CNN has better results for paying more attention to details in images,	(H. Zhang et al., 2022)
		MLP, Cond NN, MCNN and OCNN	MLP achieves an accuracy of over 96%, surpassing the 3 nearest neighbor methods.	(G KISHOR KUMAR et al., 2021a)
		CNN for handwritten characters and YOLOv4-s and YOLOv4 for printed characters.	CNN achieves an accuracy of 80.93% while YOLOv4-s boasts 99.39% accuracy, surpassing the YOLOv4 model by 20.57%.	(C. J. Lin et al., 2022)
End-to-end model		YOLOv3 + CRNN	High recognition accuracy, can identify invoice content and reduce loss.	(Yao et al., 2022)

In addition to the results summarized in the previous table, Table 3 shows the results display additional findings from the initial search that were not present in the SLR but are considered important for inclusion in the document.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 3. OCR results from the preliminary search

Task	Sub-task	Algorithms	Results	References
Character Level		Character Region Awareness for Text Detection (CRAFT)	Achieves state-of-the-art-performance	(Baek et al., 2019)

3.2.1.3 Information Extraction

In the context of this stage, the algorithms identified are presented in Table 4.

Table 4. Algorithms for IE identified in SLR

Task	Sub-task	Algorithms	Results	References
General		Proposed model uses F-RCNN and RPN. Compared with F-RCNN and YOLOv3.	Proposed model achieves an accuracy of 82.44%.	(Z. Zhang et al., 2020)
Unstructured Documents		G-CNN and LGCN; Hyperparameters of BiLSTM and MLP.	Error reduction up to 33%.	(Oral & Eryiğit, 2022)
Tables	**General**	GatedGCNs	Precision, recall, F1-Score, and accuracy are all above 0.84.	(Liu et al., 2022)
	Table detection	GNN	The best threshold is 0.1, achieving F1-Score of 73.7 with CON-ANONYM dataset, and 30.8 with RVL-CDIP dataset. It's able to deal with anonymous data because those do not use textual content.	(Riba et al., 2019)
		M-RCNN and YOLOv5	mAP was 95.14% with image processing techniques, 98.07% with YOLOv5 and 98.19% with M-RCNN, but YOLOv5 was preferred because it is faster and requires less system requirements.	(Arslan, 2022)
	Table localization	Pix2pixHD architecture	98.29% precision	(Reza et al., 2019)
	Table Segmentation	SegNet architecture	96% accuracy	(Reza et al., 2019)

Additionally to the results summarized in Table 4, the results obtained in the initial search that were not found in the SLR but are considered important for the inclusion in the document, as they involve different algorithms, are presented in Table 5.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 5. IE results from the preliminary search.

Task	Sub-task	Algorithms	Results	References
General		RNN	Outperforms Random Forest baseline	(Katti et al., 2018)

3.2.1.4 Preprocessing

In the domain of CV, the preprocessing phase plays a crucial role in improving the quality and accuracy of various applications. Several authors have emphasized the significance of preprocessing techniques, as identified in the SLR.

(Zhi et al., 2021) conducted experiments using data augmentation techniques, such as, random noise (Gaussian noise and salt-and pepper), colour jitter, horizontal lines, and random rotation. These techniques were employed to enhance the accuracy of text recognition.

Yao et al., (2022) adopted a different approach by utilizing normalization, grayscale conversion, edge detection of the original image, and the creation of binary images. These techniques effectively address issues related to poor lightening and noise effects commonly found in scanned images.

Alshathri et al., (2022) explored the use of a SDAE, SNR, PSNR, SSIM and UQI. They compared these methods with other filtering techniques, including the Nonlocal Means filter, Anisotropic diffusion filter, Gaussian filters, and Mean filters. The results clearly demonstrated the superior performance of proposed SDAE method.

3.2.2 Summary of the identified datasets

This section focuses on the 3rd RD, which pertains to the datasets utilized in the examined reports.

Invoices typically contain private data, and due to confidentiality concerns, there is a lack of publicly available datasets for the domain. However, the use of a specific invoice dataset would not be sufficient to build an effective invoice recognition system. The use of a set of validation techniques helps to determine the quality of the data to verify whether the available data is suitable for training the model (Baviskar et al., 2021). The following table, Table 6, presents the datasets that have been identified through the SLR.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Table 6. Identified datasets through the SLR

Name	Number of records	Repository	Reference
HJDataset	250.000 Japanese documents with complex layout	https://github.com/dell-research-harvard/HJDataset	(Biswas et al., 2021)
ICDAR 2013	422 images	https://paperswithcode.com/dataset/icdar-2013	(Reza et al., 2019) (Liu et al., 2022)
Pendigits	256 handwritten digits	http://archive.ics.uci.edu/dataset/81/pen+based+recognition+of+handwritten+digits	(G KISHOR KUMAR et al., 2021)
PubLayNet	358.353 document images	https://github.com/ibm-aur-nlp/PubLayNet	(Biswas et al., 2021)
RVL-CDIP	400.000 scanned document images	https://paperswithcode.com/dataset/rvl-cdip	(Riba et al., 2019)
SciTSR	15.000 tables	https://github.com/Academic-Hammer/SciTSR	(Liu et al., 2022)
TableBank	417.234 tables	https://github.com/doc-analysis/TableBank	(Lee & Chen, 2021)
Table-detection-dataset	400 table file images	https://github.com/sgrpanchal31/table-detection-dataset	(Lee & Chen, 2021)

In addition to the datasets identified in the SLR, an extra search was carried out to locate datasets commonly used for various processes outlined in the CV approaches.

Table 7. Other datasets identified.

Name	Number of records	Repository	References
CIFAR-10	60.000 images	https://www.cs.toronto.edu/~kriz/cifar.html	(Suganuma et al., 2017)
COCO-Text	22.184 images	https://bgshih.github.io/cocotext/	(Veit et al., 2016)
CTW1500	10.751 images	https://github.com/Yuliang-Liu/Curve-Text-Detector	(Subramani et al., 2020)
CUTE80	80 images	http://cs-chan.com/downloads_cute80_dataset.html	(Atienza, 2021)
DocBank	500.000 document images	https://doc-analysis.github.io/docbank-page/ https://github.com/doc-analysis/DocBank	(Li et al., 2020)
FUNSD	199 noisy scanned documents	https://guillaumejaume.github.io/FUNSD/download/ https://huggingface.co/datasets/nielsr/FUNSD_layoutlmv2	(Jaume et al., 2019)
ICDAR 2019	10.166 images	https://github.com/cndplab-founder/ICDAR2019_cTDaR	(Ha & Horák, 2022); (X. Zhao et al., 2019)
ImageNet	14.197.122 images	https://www.kaggle.com/competitions/imagenet-object-localization-challenge/data	(Studer et al., 2019)
Marmot	7.907 document pages	https://www.icst.pku.edu.cn/cpdp/sjzy/	(Li et al., 2019)
Microsoft COCO	328.000 images	https://cocodataset.org/#download	(Zou et al., 2019)
MMOCR	images	https://mmocr.readthedocs.io/en/dev-1.x/user_guides/dataset_prepare.html	(Huang et al., 2021)
MSRA-TD500	500 images	http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500)	(Risnumawan et al., 2014)

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Open Images	+10.000.000 images	https://storage.googleapis.com/openimages/web/download_v7.html	(Zou et al., 2019)
Pascal VOC	2.913 images	https://pjreddie.com/projects/pascal-voc-dataset-mirror/	(Zou et al., 2019)
RCTW-17	12.000 images	https://rctw.vlrlab.net/dataset	(Huang et al., 2021)
SynthText	800.000 images	https://github.com/ankush-me/SynthText	(Gupta et al., 2016)
TableSence	2.615 tables	https://github.com/microsoft/TableSense	(Dong et al., 2019)
Total-Text	1.555 images	https://www.kaggle.com/datasets/konradb/text-recognition-total-text-dataset https://github.com/cs-chan/Total-Text-Dataset	(Huang et al., 2021)

Another intriguing perspective on the datasets utilized in the literature is to associate them by the specific task and sub-task they address. This approach allows for a more granular analysis, enabling researchers to discern how different datasets are strategically chosen to align with a particular aspect of the research, fostering a deeper understanding of the interplay between datasets and the objectives they serve.

Table 8. Dataset for each process and their task and sub-tasks

Process stage	Task	Sub-task	Datasets
Preprocessing			CIFAR-10
LA	**General**		DocBank; Marmot; HJDataset; RVL-CDIP; PubLayNet;
OCR	**General**		COCO-Text; CTW1500; CUTE80; DocBank; FUNSD ICDAR 2013; Microsoft COCO; MSRA-TD500; Pascal VOC; SynthText; Total-Text
	Text Detection	**As object detection**	ImageNet; Open Images
	Text Recognition		Pendigits
	End-to-end model		MMOCR; RCTW-17
IE	**Table**	**General**	ICDAR 2019; SciTSR; TableBank
		Table Detection	TableSence; Table-detection-dataset
		Table Recognition	Marmot

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

4 DISCUSSION

In the discussion section, we will delve into the crucial aspects of invoice recognition based on the analysis of results from the reports. The discussion is about responding to the questions previously selected, and responding based on the results of the reports, offering valuable insights into each of the key components of invoice recognition and their significance in the overall process.

4.1 RQ1: What are the essential stages involved in conducting an invoice recognition process?

After a comprehensive analysis of the reports, four primary stages have been identified as essential components for conducting an effective invoice recognition process, such as image preprocessing, LA, OCR, and IE. These stages play a critical role in accurately and efficiently extracting information from invoices. It's important to note that while all four stages are not always required simultaneously, their combination is often necessary to achieve optimal results.

4.1.1 RQ1.1: Why is it important to preprocess images?

The initial stage in the invoice recognition process involves image preprocessing. During this stage, the raw images are prepared for further analysis, this may techniques tasks such as noise reduction, image enhancement, and correction of the angle, but also data augmentation that enhance the model's ability to handle real-world variation and improve its generalization.

For invoice recognition, introducing controlled noise, as shown in Figure 5, or rotation, as shown in Figure 6, to training images can be valuable to create a more robust model that can deal with various noise patters. This is particularly critical since the invoices subjected to recognition processes are often scanned, which may result in a loss of image quality.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

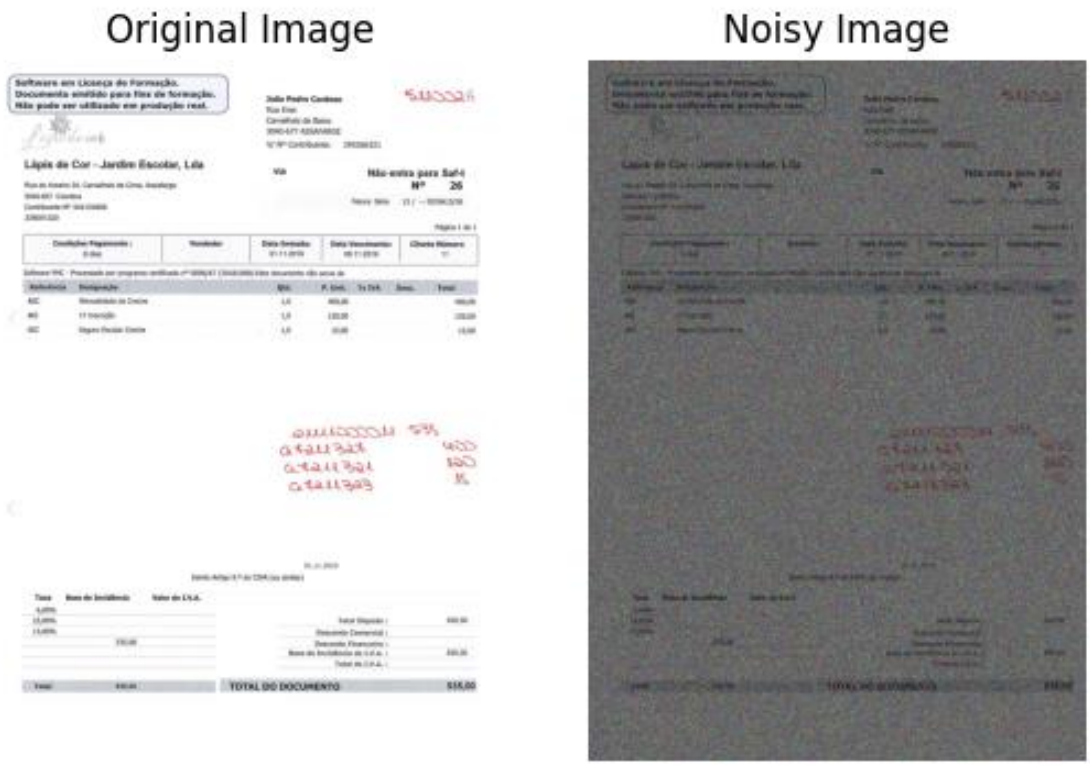

Figure 5. Adding noise to images

Figure 6. Adding rotation to images

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

In Yao et al., (2022) and (Zhi et al., 2021) studies, image processing was used on raw images before analysis, as a result, solves poor lightening and noisy effect of scanned images and improved the accuracy of text detection, respectively.

4.1.2 RQ1.2: How does LA improve invoice recognition?

LA is the process of understanding the structure and organization of invoice documents, guiding the extraction of valuable information. It plays a crucial role in the overall recognition process by identifying and delineating essential regions within an invoice, including headers, footers, and tables, and other key components. These segmentation process, illustrated in Figure 7, enables the isolation and separate processing of these regions in subsequent stages.

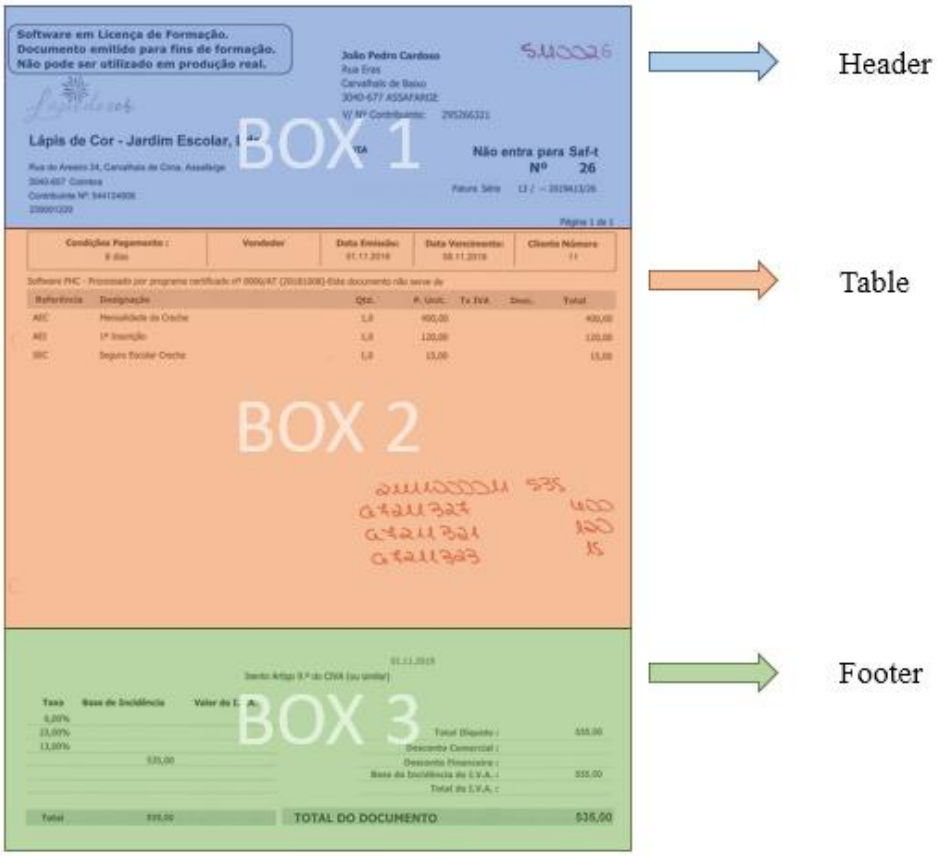

Figure 7. Example of a Document Layout Analysis

The application of LA offers several advantages in the context of invoice recognition. By dividing the document into meaningful regions, LA streamlines the subsequent OCR and

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

IE stages. It aids in precise text extraction from specific areas, resulting in more accurate and efficient recognition of invoice details.

4.2 RQ2: Which algorithms are commonly employed to execute distinct stages, as LA, OCR, and IE?

In the field of LA, Biswas et al., (2021) stood out by employing a model inspired by M-RCNN algorithms and incorporating the ResNeXt-101 architecture. This innovative approach resulted in an impressive mAP of 0.904, showcasing the model's remarkable efficiency in identifying and analysing layout structures. The proposed method demonstrated an AP exceeding 0.820, underscoring the importance of algorithmic choices in LA.

Moving to OCR, S. Zhao et al. (2020) and H. Zhang et al. (2022) utilized state-of-the-art techniques. MobileNet proved to be an excellent choice for text detection, significantly reducing time consumption while enhancing hardware processes. Additionally, the PAN-640 architecture excelled in terms of accuracy and speed, highlighting the practicality of these algorithms in the OCR domain.

In the domain of Text Recognition, the CRNN algorithm, as implemented by (Zhi et al., 2021), delivered outstanding results. It achieved an impressive accuracy rate of 99.03% with image enhancement, all while maintaining minimal time consumption when utilizing image slicing. These findings demonstrate the significance of algorithm selection in optimizing text recognition.

For character-level tasks, C. J. Lin et al. (2022) found that the YOLOv4-s algorithm for printed characters surpassed the conventional YOLOv4 model by a remarkable 20.57% in terms of accuracy. These results underscore the critical role of algorithm selection in character recognition tasks.

In the realm of IE, Z. Zhang et al. (2020) presented a model based on F-RCNN and RPN, achieving an impressive accuracy rate of 82.44%. This result exemplifies the model's proficiency in identifying and extracting valuable information in the context of general IE tasks.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

For unstructured documents, Oral & Eryiğit (2022) employed G-CNN and LGCN with optimized hyperparameters. This approach led to a substantial reduction in error of up to 33%, emphasizing the significance of algorithm selection in improving information extraction from diverse and unstructured documents.

In the domain of table IE, Liu et al. (2022) found that GatedGCN delivered exceptional results. They achieved precision, recall, F1-Score, and accuracy metrics, all surpassing 0.84. These findings highlight the importance of algorithmic choices in optimizing table IE.

These exemplary results provide critical insights into the prowess of various algorithms and architectures used in different stages of invoice recognition, emphasizing their roles in achieving remarkable accuracy and efficiency. These achievements further advance the field, bringing us close to a robust and effective invoice recognition system.

4.3 RQ3: What software and hardware configurations are commonly used in these CV approaches?

It's clear from the reports that many of researchers are utilizing high-performance GPUs in their configurations, NVIDIA GPU's, such as the GTX1080, Tesla V100, and GeForce GTX 1080 Ti, are common choices. These GPUs are known for their capabilities in accelerating DL tasks and are often favored for CV applications. High CPU capabilities, such as Intel Core i7, are also mentioned, which can be useful for preprocessing and handling data.

Several reports specify ample RAM, ranging from 4GB to 64GB. Having sufficient RAM is crucial for handling large datasets and complex DL models, ensuring smoother and more efficient processing.

The software configurations vary significantly among the reports. Researchers often use combinations of open-source libraries like TensorFlow, PyTorch, and OpenCV. The choice of software stack may depend on factors such as familiarity with the tools and specific requirements of the research.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Reports mention the use of different operating systems, including various versions of Windows, Linus (CentOS, Ubuntu), the choice can be influenced by the specific needs of the research and the compatibility of the tools being used.

Overall, the reports reflect a high standard of software and hardware quality. Researchers tend to invest in robust hardware configurations, especially powerful GPUs and substantial memory, to ensure the efficiency and accuracy of their CV models. The diversity of software tools reflects the flexibility of their field, allowing researchers to choose the best tools for their specific research objectives. These common practices are indicative of a commitment to quality and efficiency within the field of CV for invoice recognition.

4.4 RQ4: Which programming languages are frequently utilized in the development of invoice recognition process?

Based on the analysis of the reports provided, Python appears to be the most frequently utilized programming language in the development of invoice recognition processes.

4.5 RQ5: Which datasets are typically employed for research purposes within each stage of invoice recognition process?

The fact that many of the datasets used in the reports are unavailable due to confidentiality or privacy reasons is a common challenge in the field of invoice recognition research. This challenge underscores the sensitive nature of financial and transactional documents like invoices. Researchers must respect the privacy associated with these documents, which restricts their ability to openly shave or use publicly available datasets.

In this context, the use of datasets beyond just invoice documents can have a positive impact on improving the performance and versability of models for invoice recognition, such as PubLayNet and ImageNet. This approach allows researchers to leverage knowledge and techniques developed in related domain, contributing to the robustness and generalization of invoice recognition models.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

4.6 Key Findings

Our analysis revealed the essential stages in invoice recognition process, emphasizing the importance of image preprocessing, LA, OCR, and IE. Image preprocessing stands out as a key element for enhancing model robustness, particularly through techniques like noise reduction and data augmentation. LA aids in accurately identifying invoice elements and the information they contain, facilitating the application of OCR and IE.

Additionally, our examination underscores the significance of software and hardware configurations, with powerful GPUs and optimizing software frameworks contribute to overall process efficiency.

4.7 Limitations

One notable limitation in the field of invoice recognition is the scarcity of publicly available datasets, often due to confidentiality constrains. However, this challenge can be partially mitigated by leveraging publicly available datasets that are specific to certain elements of invoices, such as tables.

Another limitation that warrants considerations is the substantial computational resources required for efficient invoice recognition. This includes the need for powerful GPU's and CPUs, which may pose practical constraints for smaller research groups with limited resources. Addressing this limitation could involve the development of more efficient algorithms, making the invoice recognition process more accessible and cost-effective.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

5 CONCLUSION

After a comprehensive analysis of the selected reports, this study has delved into the realm of invoice recognition using AI. The investigation aimed to identify key stages, algorithms, software and hardware configurations, programming languages, and datasets commonly employed in this domain. The research also addresses the importance of image preprocessing, LA, and significance of various algorithm choices, shedding light on the challenges faced within this field.

The findings of this research have made several important contributions to the field of invoice recognition. The essential stages involved in the recognition process, including image preprocessing, LA, OCR, and IE, were clearly delineated, underlining their crucial roles in achieving accurate and efficient results. Notably, image preprocessing emerged as a pivotal stage for enhancing model robustness through techniques like noise reduction and data augmentation. LA was identified as a fundamental stage for streamlining subsequent OCR and IE, resulting in more precise recognition. Furthermore, this study highlighted the importance of algorithm selected in each stage, offering insights into the specific algorithm and architectures that have excelled in recent research. The significance of powerful GPUs, ample memory, and well-optimized process was underscored. Python was identified as the predominant programming language in this domain, while acknowledging the diversity of software tools and hardware configurations. Finally, this research recognized the challenges posed by the limited availability of publicly accessible datasets, particularly due to confidentiality constraints. It suggested potential solutions, such as leveraging related domain datasets, to overcome these limitations.

While this research has advanced the understanding of invoice recognition processes, it also revealed certain limitations in the field. The scarcity of publicly available datasets, primarily due to confidentiality constraints, remains a significant challenge. Privacy concerns surrounding financial and transactional documents like invoices limit researchers ability to openly share or use such data. Furthermore, the substantial computational resources required for efficient invoice recognition, including powerful GPUs and CPUs, may pose practical constraints, particularly for smaller research groups

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

with limited resource. Addressing this limitation could involve developing more efficient algorithms to make the recognition process more accessible and cost-effective.

Looking ahead, this study serves as a foundational step in the realm of AI for invoice recognition. As technology and research continue to advance, numerous promising areas for future exploration and development emerge.

Researchers in the field of invoice recognition can explore various avenues to address the limitations of available datasets. Collaboration with organizational willing to shake anonymized data may prove instrumental in expanding and diversifying datasets for improved model training and testing. Additionally, efforts aimed at developing more efficient algorithms that reduce the computational resource demands are essential to make this field more accessible and cost-effective. Furthermore, continuous exploration of algorithm choices, including the investigation of novel architectures and approaches, holds great potential for enhancing accuracy and efficiency.

In conclusion, this study provides a solid foundation for further advancements in the domain of AI for invoice recognition. It illuminates potential directions for future research and development, offering a roadmap to tackle existing challenge and leverage new opportunities in this evolving field.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

REFERENCES

Adam Hayes. (2023, September 27). *What Is an Invoice? It's Parts and Why They Are Important*. Investopedia. https://www.investopedia.com/terms/i/invoice.asp#:~:text=An%20invoice%20is%20a%20document%20that%20maintains%20a,must%20be%20approved%20by%20the%20responsible%20management%20personnel.

Aladhadh, S., Rehman, H. U., Qamar, A. M., & Khan, R. U. (2021). Recurrent convolutional neural network MSER-based approach for payable document processing. *Computers, Materials and Continua*, *69*(3), 3399–3410. https://doi.org/10.32604/cmc.2021.018724

Alshathri, S. I., Vincent, D. J., & Hari, V. S. (2022). Denoising letter images from scanned invoices using stacked autoencoders. *Computers, Materials and Continua*, *71*(1), 1371–1386. https://doi.org/10.32604/cmc.2022.022458

Arslan, H. (2022). End to End Invoice Processing Application Based on Key Fields Extraction. *IEEE Access*, *10*, 78398–78413. https://doi.org/10.1109/ACCESS.2022.3192828

Atienza, R. (2021). *Data Augmentation for Scene Text Recognition*. http://arxiv.org/abs/2108.06949

Baek, Y., Lee, B., Han, D., Yun, S., & Lee, H. (2019). *Character Region Awareness for Text Detection*.

Baviskar, Di., Ahirrao, S., & Kotecha, K. (2021). Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches. *IEEE Access*, *9*, 101494–101512. https://doi.org/10.1109/ACCESS.2021.3096739

Biswas, S., Riba, P., Lladós, J., & Pal, U. (2021). Beyond document object detection: instance-level segmentation of complex layouts. *International Journal on Document Analysis and Recognition*, *24*(3), 269–281. https://doi.org/10.1007/s10032-021-00380-6

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Cardie, C. (1997). *Empirical Methods in Information Extraction) (© AAAI)* (Vol. 18). www.junglee.com/suc-

Chi, Z., Huang, H., Xu, H.-D., Yu, H., Yin, W., & Mao, X.-L. (2019). *Complicated Table Structure Recognition*.

Dong, H., Liu, S., Han, S., Fu, Z., & Zhang, D. (2019). *TableSense: Spreadsheet Table Detection with Convolutional Neural Networks*. www.aaai.org

Dua, D., & Graff, C. (2019). UCI Machine Learning Repository. *University of California, School of Information and Computer Science, Irvine, CA.*

G KISHOR KUMAR, R RAJA KUMAR, RAM CHAKKA, & P VISWANATH. (2021). *A multi-pronged accurate approach to optical character recognition, using nearest neighborhood and neural-network-based principles*.

Gao, L., Huang, Y., Dejean, H., Meunier, J.-L., Yan, Q., Fang, Y., Kleber, F., & Lang, E. (2019). ICDAR 2019 Competition on Table Detection and Recognition (cTDaR). *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1510–1515. https://doi.org/10.1109/ICDAR.2019.00243

Gobel, M., Hassan, T., Oro, E., & Orsi, G. (2013). ICDAR 2013 Table Competition. *2013 12th International Conference on Document Analysis and Recognition*, 1449–1453. https://doi.org/10.1109/ICDAR.2013.292

Gupta, A., Vedaldi, A., & Zisserman, A. (2016). *Synthetic Data for Text Localisation in Natural Images*. http://arxiv.org/abs/1604.06646

Ha, H. T., & Horák, A. (2022). Information extraction from scanned invoice images using text analysis and layout features. *Signal Processing: Image Communication*, *102*. https://doi.org/10.1016/j.image.2021.116601

Harley, A. W., Ufkes, A., & Derpanis, K. G. (2015). Evaluation of deep convolutional nets for document image classification and retrieval. *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 991–995. https://doi.org/10.1109/ICDAR.2015.7333910

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Holeček, M. (2020). *Learning from similarity and information extraction from structured documents*. https://doi.org/10.1007/s10032-021-00375-3

Huang, J., Pang, G., Kovvuri, R., Toh, M., Liang, K. J., Krishnan, P., Yin, X., Hassner, T., & Ai, F. (2021). *A Multiplexed Network for End-to-End, Multilingual OCR*.

Jaume, G., Ekenel, H. K., & Thiran, J.-P. (2019). *FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents*. http://arxiv.org/abs/1905.13538

Katti, A. R., Hoehne, J., Bickel, S., & Faddoul, J. B. (2018). *Applying Sequence-to-Mask Models for Information Extraction from Invoices*. https://blog.altoros.com/optical-character-

Lee, S. H., & Chen, H. C. (2021). U-ssd: Improved ssd based on u-net architecture for end-to-end table detection in document images. *Applied Sciences (Switzerland)*, *11*(23). https://doi.org/10.3390/app112311446

Li, M., Cui, L., Huang, S., Wei, F., Zhou, M., & Li, Z. (2019). *TableBank: A Benchmark Dataset for Table Detection and Recognition*.

Li, M., Xu, Y., Cui, L., Huang, S., Wei, F., Li, Z., & Zhou, M. (2020). *DocBank: A Benchmark Dataset for Document Layout Analysis*. http://arxiv.org/abs/2006.01038

Lin, C. J., Liu, Y. C., & Lee, C. L. (2022). Automatic Receipt Recognition System Based on Artificial Intelligence Technology. *Applied Sciences (Switzerland)*, *12*(2). https://doi.org/10.3390/app12020853

Lin, Y., Ji, H., Huang, F., & Wu, L. (2020). *A Joint Neural Model for Information Extraction with Global Features*. http://blender.cs.illinois.edu/software/

Liu, Y., Liang, X., Chen, S., Diao, L., Tang, X., Fang, R., & Chen, W. (2022). Table information extraction and analysis: A robust geometric approach based on GatedGCN. *Proceedings - International Conference on Pattern Recognition*, *2022-August*, 3131–3137. https://doi.org/10.1109/ICPR56361.2022.9956139

Liu, Y., Soh, L.-K., & Lorang, E. (2021). Investigating coupling preprocessing with shallow and deep convolutional neural networks in document image classification. *Journal of Electronic Imaging*, *30*(04). https://doi.org/10.1117/1.jei.30.4.043024

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Lopez, M. M., & Kalita, J. (2017). *Deep Learning applied to NLP*. http://arxiv.org/abs/1703.03091

Mohsin Reza, M., Ajraf Rakib, M., Saqib Bukhari, S., & Dengel, A. (2018). *A High-Performance Document Image Layout Analysis for Invoices*.

Novais, J. (2023). *Online Knowledge Library*. Https://Www.b-on.Pt/En/What-Is-b-On/. https://fccn.pt/en/conhecimento/b-on/

Oral, B., & Eryiğit, G. (2022). Fusion of visual representations for multimodal information extraction from unstructured transactional documents. *International Journal on Document Analysis and Recognition*, *25*(3), 187–205. https://doi.org/10.1007/s10032-022-00399-3

Özgen, A. C., Fasounaki, M., & Ekenel, H. K. (2018). Text detection in natural and computer-generated images. *26th IEEE Signal Processing and Communications Applications Conference, SIU 2018*, 1–4. https://doi.org/10.1109/SIU.2018.8404600

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. In *The BMJ* (Vol. 372). BMJ Publishing Group. https://doi.org/10.1136/bmj.n71

Palm, R. B., Winther, O., & Laws, F. (2017). *CloudScan - A configuration-free invoice analysis system using recurrent neural networks*. http://arxiv.org/abs/1708.07403

Rashid, S. F., Akmal, A., Adnan, M., Aslam, A. A., & Dengel, A. (2017). Table Recognition in Heterogeneous Documents Using Machine Learning. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, *1*, 777–782. https://doi.org/10.1109/ICDAR.2017.132

Reza, M. M., Bukhari, S. S., Jenckel, M., & Dengel, A. (2019). Table localization and segmentation using GAN and CNN. *2019 International Conference on Document*

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Analysis and Recognition Workshops, ICDARW 2019, 152–157. https://doi.org/10.1109/ICDARW.2019.40097

Riba, P., Dutta, A., Goldmann, L., Fornés, A., Ramos, O., & Lladós, J. (2019). *Table Detection in Invoice Documents by Graph Neural Networks*. https://zenodo.org/record/3257319

Risnumawan, A., Shivakumara, P., Chan, C. S., & Tan, C. L. (2014). A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*, *41*(18), 8027–8048. https://doi.org/10.1016/j.eswa.2014.07.008

Schreiber, S., Agne, S., Wolf, I., Dengel, A., & Ahmed, S. (2017). DeepDeSRT: Deep Learning for Detection and Structure Recognition of Tables in Document Images. *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR*, *1*, 1162–1167. https://doi.org/10.1109/ICDAR.2017.192

Shen, Z., Zhang, K., & Dell, M. (2020). *A Large Dataset of Historical Japanese Documents with Complex Layouts*.

Studer, L., Alberti, M., Pondenkandath, V., Goktepe, P., Kolonko, T., Fischer, A., Liwicki, M., & Ingold, R. (2019). *A Comprehensive Study of ImageNet Pre-Training for Historical Document Image Analysis*.

Subramani, N., Matton, A., Greaves, M., & Lam, A. (2020). *A Survey of Deep Learning Approaches for OCR and Document Understanding*. http://arxiv.org/abs/2011.13534

Suganuma, M., Shirakawa, S., & Nagao, T. (2017). A genetic programming approach to designing convolutional neural network architectures. *GECCO 2017 - Proceedings of the 2017 Genetic and Evolutionary Computation Conference*, 497–504. https://doi.org/10.1145/3071178.3071229

Susheela Devi, V., & Murty, M. N. (2002). An incremental prototype set building technique. *Pattern Recognition*, *35*(2), 505–513. https://doi.org/10.1016/S0031-3203(00)00184-9

Table-Detection-Dataset. (2023). https://github.com/sgrpanchal31/table-detection-dataset

Veit, A., Matera, T., Neumann, L., Matas, J., & Belongie, S. (2016). *COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images*. http://arxiv.org/abs/1601.07140

Wenkel, S., Alhazmi, K., Liiv, T., Alrshoud, S., & Simon, M. (2021). Confidence score: The forgotten dimension of object detection performance evaluation. *Sensors*, *21*(13). https://doi.org/10.3390/s21134350

Yao, X., Sun, H., Li, S., & Lu, W. (2022). Invoice Detection and Recognition System Based on Deep Learning. *Security and Communication Networks*, *2022*. https://doi.org/10.1155/2022/8032726

Zhang, H., Dong, B., Zheng, Q., & Feng, B. (2022). Research on fast text recognition method for financial ticket image. *Applied Intelligence*, *52*(15), 18156–18166. https://doi.org/10.1007/s10489-022-03467-7

Zhang, H., Dong, B., Zheng, Q., Feng, B., Xu, B., & Wu, H. (2022). All-content text recognition method for financial ticket images. *Multimedia Tools and Applications*, *81*(20), 28327–28346. https://doi.org/10.1007/s11042-022-12741-2

Zhang, Z., Zhang, D., Jin, T., & Zhang, M. (2020). Form location and extraction based on deep learning. *IOP Conference Series: Materials Science and Engineering*, *768*(5). https://doi.org/10.1088/1757-899X/768/5/052082

Zhao, S., Sun, L., Li, G., Liu, Y., & Liu, B. (2020). A CCD based machine vision system for real-time text detection. *Frontiers of Optoelectronics*, *13*(4), 418–424. https://doi.org/10.1007/s12200-019-0854-0

Zhao, X., Niu, E., Wu, Z., & Wang, X. (2019). *CUTIE: Learning to Understand Documents with Convolutional Universal Text Information Extractor*. http://arxiv.org/abs/1903.12363

Zhao, Z. Q., Zheng, P., Xu, S. T., & Wu, X. (2019). Object Detection with Deep Learning: A Review. In *IEEE Transactions on Neural Networks and Learning Systems* (Vol. 30, Issue 11, pp. 3212–3232). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/TNNLS.2018.2876865

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

Zhi, X., Shen, Z., & Zhao, B. (2021). A method for identifying the key information of electronic invoicing in complex scenes. *2021 6th International Conference on Image, Vision and Computing, ICIVC 2021*, 90–94. https://doi.org/10.1109/ICIVC52351.2021.9526973

Zhong, X., Tang, J., & Yepes, A. J. (2019). *PubLayNet: largest dataset ever for document layout analysis*.

Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). *Object Detection in 20 Years: A Survey*. http://arxiv.org/abs/1905.05055

Zucker, A., Belkada, Y., Vu, H., & Nguyen, V. N. (2021). ClusTi: Clustering Method for Table Structure Recognition in Scanned Images. *Mobile Networks and Applications*, *26*(4), 1765–1776. https://doi.org/10.1007/s11036-021-01759-9

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIXS

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 1. Metrics for classification models

To evaluate the results of the models, in Table 9 are presented some evaluation metrics.

Table 9 Metric for Classification Models

Metric	Description
Precision	Measures the reliability of the extracted data (Cardie, 1997).
Intersection over Union (IoU)	Calculates how much a candidate boundary block overlaps the true boundary block (intersection), divided by the total space occupied by the candidate block and the true block (union) (Subramani et al., 2020).
F1-Score	Calculated to evaluate the performance of the model (Subramani et al., 2020).
Recall	Measures the amount of relevant information extracted correctly (Cardie, 1997).
AP	Overall performance, measured in terms of accuracy per class, where a class is determined to be correct only when overy token in the class is correct (X. Zhao et al., 2019).
softAP	Indicates the ability of the model to extract correct key information with tolerance for false positives (X. Zhao et al., 2019).
Confidence score	Eliminates false positives and ensures that a predicted bounding box has a certain minimum score (Wenkel et al., 2021).
Levenshtein distance	Measures the distance between two sequences, namely the true information and the extracted information (Ha & Horák, 2022).

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

APPENDIX 2. Articles from SLR not used in the results

ClusTi - Clustering Method for Table Structure Recognition in Scanned Images

Zucker et al., (2021) proposes an efficient method called CluSTi (Clustering method for recognition of the Structure of Tables in invoice scanned Images). The contributions of CluSTi are three-fold. Firstly, it removes heavy noises in the table images using a clustering algorithm. Secondly, it extracts all text boxes using state-of-the-art text recognition. Thirdly, based on the horizontal and vertical clustering algorithm with optimized parameters, CluSTi groups the text boxes into their correct rows and columns, respectively.

- Algorithms: Character Region Awareness for Text Detection (CRAFT) and Desteny-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithms.
- Hardware/Software Configuration: Not specified.
- Programming Language: Python; Code: Not available.
- Dataset(s): Public scanned images (not available), ICDAR 2013 (Gobel et al., 2013) and ICDAR 2019 (Gao et al., 2019).

The experiment results demonstrate the effectiveness of the CluSTi approach in recognizing table structures in scanned invoice images. CluSTi achieved an F1-score of 87.5%, 98.5%, and 94.5%, for scanned images, ICDAR 2013 and ICDAR 2019, respectively. Notably, the proposed method outperformed DeepDeSRT with an F1-score of 98.48% compared to 91.44%.

Investigating coupling preprocessing with shallow and deep convolutional neural networks in document image classification

The study conducted by Liu et al., (2021) is focused on understanding the influence of preprocessing techniques on the performance of CNNs concerning their effectiveness and efficiency. This research encompasses an investigation into various aspects of CNN performance, particularly in the context of classifying noisy printed document images, including historical newspapers.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

- Algorithms: CNN, and architectures LeNet-5, LeNet-7, LeNet-9, ResNet-18, ResNet-152, MobileNetV2, and EfficientNet.
- Hardware/Software Configuration: Not specified.
- Programming Language: Code: Not Available.
- Dataset: RVL-CDIP (Harley et al., 2015).

Preprocessing, such as Light-Otsu, Light-Howe, Aggressive-Otsu and Aggressive-Howe, significantly enhances CNN performance, particularly in situations with limited training data. Deep CNNs coupled with preprocessing can outperform Very Deep CNNs effectively and efficiently; and aggressive preprocessing is not helpful as it could remove potentially useful information in document images. CNN models coupled with preprocessing could outperform those without preprocessing in cases in which there were smaller training samples.

Learning from similarity and information extraction from structured documents

Holeček, (2020) expand previous work where was proved that convolutions, graph convolutions and self-attention can work together and exploit all the information present in a structured document, and design and examine various approaches to using Siamese networks, concepts of similarity, one-shot learning, and context and memory awareness.

- Algorithms: CNN and GCN.
- Hardware/Software Configuration: Open-source library: TensorFlow
- Programming Language: Python; Code: https://github.com/Darthholi/similarity-models
- Dataset: Not available.

The results verify the hypothesis that trainable access to a similar yet distinct page, in conjunction with pre-existing target information, profoundly enhances the process of information extraction from structured documents. Furthermore, the series of experiments conducted throughout the study consistently substantiate that all proposed architectural components are indispensable, including Siamese networks, employing class information, query-answer attention module and skip connections to a similar page, all

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt

are integral to surpassing the performance benchmarks established by previous methodologies.

The best model improves the previous state-of-the art results by an 8.25 % gain in F1-score. This substantial performance boost translates into tangible benefits, estimated at approximately $4.000 in monthly cost savings for a medium-sized enterprise. Qualitative analysis is provided to verify that the new model performs better for all target classes.

U-SSD: Improved SSD Based on U-Net Architecture for End-to-End Table Detection in Document Images

Due to the different arrangements of tables and texts, as well as the variety of layouts, table detection is a challenge in the field of document analysis. Therefor (Lee & Chen, 2021) proposes an end-to-end table detection model, U-SSD, as based on the object detection method of DL, takes the Single Shot MultiBox Detector (SSD) as the basic model architecture, improves it by U-Net, and adds dilated convolution to enhance the feature learning capability of the network.

- Algorithms: SSD improved with U-Net architectures. Compared with Faster R-CNN+VGG, Faster R-CNN+ResNet50, YOLOv3, YOLOv3+U-Net and U-SSD.
- Hardware/Software Configurations: CPU: Intel Core i7-9700; CPU: 3.00 GHz*8; GPU: GeForce RTX 2080 Ti/PCIe/SSE2; OS: Linux Ubuntu 18.04.3
- Deep Learning Framework: PyTorch.
- Programming Language: Python; Code: Not available.
- Dataset(s): Tabular data provided by a Taiwanese Law Firm (not available), TableBank (Li et al., 2019), Github open datasets (*Table-Detection-Dataset*, 2023) and ICDAR13 (Gobel et al., 2013).

The experimental results show that the proposed method is effective. The improved SSD-based network architectures can further improve the accuracy of table detection and minimize the prediction error at table edges, and the public dataset verification results show that the detection effect is good. The F1-score of SSD is higher than 0.80, reaching 0.94 when adding Dilatation and U-Net. Using an image segmentation model for object detection can also achieve good results and adding dilated convolution can effectively improve feature information.

Coimbra Business School | ISCAC
Quinta Agrícola – Bencanta, 3045-601 Coimbra
Tel +351 239 802 000; E-mail: secretariado@iscac.pt; www.iscac.pt