

# INVOICE DATA EXTRACTION

Documented work on 11th June

## 1. Deep Learning Model - InvoiceNet - YOLOv3 +CRNN

In order to meet the requirements of efficiently identifying invoice data in engineering applications, this paper first uses the YOLOv3 algorithm for text target detection training. Second, the deep learning CRNN model is used to identify the content of the invoice. Finally, the two models are combined to obtain an end-to-end invoice recognition model, which is verified by the test set, and the recognition result is compared with the recognition result of the traditional OCR technology.

It locates the invoice information area by marking the invoice dataset and realises the detection and recognition of the VAT invoice information through image processing and deep learning. Finally, the system realised the rapid identification and processing of invoices.


Code :

<https://github.com/naiveHobo/InvoiceNet>

Hindawi  
Security and Communication Networks  
Volume 2022, Article ID 8032726, 10 pages  
<https://doi.org/10.1155/2022/8032726>

### Research Article

### Invoice Detection and Recognition System Based on Deep Learning

Xunfeng Yao , Hao Sun, Sijun Li, and Weichao Lu

Jinling College, Nanjing University, Nanjing, China

Correspondence should be addressed to Xunfeng Yao; 030504@jlxj.nju.edu.cn

Received 13 August 2021; Accepted 29 September 2021; Published 25 January 2022

Academic Editor: Xuyun Zhang

Copyright © 2022 Xunfeng Yao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the development of economy and information technology, a large amount of invoice information has been produced. As one of the important components of the industrial Internet of Things, 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 low efficiency in identifying invoice information, error-prone, and difficulty in ensuring security frequently appear. In response to the above problems, this paper designs and implements an invoice information recognition system based on deep learning. The system first solves the problems of low image contrast and lack of image due to poor lighting or noise effects by image preprocessing methods such as image graying and normalization. Second, a target detection and invoice recognition method based on the combination of YOLOv3 + CRNN two models is proposed, and an end-to-end invoice information recognition model is obtained. Finally, the model is used to develop an invoice detection and recognition system based on deep learning. Experiments have verified that the system has the characteristics of high recognition accuracy and high efficiency, which can accurately identify invoice content information and reduce the loss of manpower and material resources.

### 1. Introduction

Foreign research on invoice recognition system originated in the 1960s and 1970s. But, most research is only on methods of invoice recognition and digital recognition. The concept of OCR (Optical Character Recognition) technology was first proposed by German scientist Tauschek in 1929. As an important part of pattern recognition, OCR is used to identify the information in the image and extract it into computer readable [1]. Until about the 1960s, Japan began to study the basic recognition theory of OCR. After more than ten years of research, it developed a simple recognition system such as postal code recognition, which realized the automatic recognition of codes on mails [2]. After 1970, China began to study OCR technology and first carried out relevant research on Chinese character recognition. Until 1986, Tsinghua University and other universities developed an invoice recognition system based on OCR technology, and Chinese OCR invoice recognition products came out [3]. Due to the low recognition rate of the early invoice

system and insufficient productization, it has not been popularized in life. With the rise of artificial intelligence, more systems for invoice recognition have begun to appear on the market. For example, Baidu's OCR recognition system and Tencent's OCR recognition system both use in-depth learning to detect and recognize invoice information [4].

Once deep learning has emerged, it has been widely used in speech recognition, image recognition, and natural language processing. In 2011, Google applied deep learning to speech recognition and successfully reduced the error rate [5, 6]. In the field of image recognition, researchers have further proposed a large-scale deep convolutional neural network, which reduces the error detection rate to 15.3% [7]. In 2015, He et al. proposed the ResNet architecture to improve the accuracy of the algorithm by increasing the amount of data during training [8]. Deep learning has developed rapidly in image recognition, and target detection technology has been applied to text localization in natural scenes. Girshick et al. proposed that R-CNN successfully

## **OTHER GITHUB REPOS:**

<https://github.com/abhayhk2001/invoice-data-extraction>

<https://github.com/invoice-x/invoice2data>

<https://github.com/piyushmathur17/invoice-extractor>

## **API BASED SOLUTIONS :**

### **1. Document Intelligence:**

API: The Document Intelligence invoice model (prebuilt-invoice) extracts key information from invoices, including customer name, billing address, due date, and amount due .

Studio: The Document Intelligence Studio allows users to analyze invoices and extract data using various tools and libraries, including C# SDK, Python SDK, Java SDK, and JavaScript SDK

### **2. DocuClipper:**

Software: DocuClipper is a specialized web-based tool that converts invoices into structured data formats like Excel, CSV, and PDF, simplifying accounts payable processes[2].

### **3. Mindee:**

API: Mindee's invoice OCR API extracts data from invoices, including customer and supplier information, amounts, and invoice identifiers, with high accuracy and real-time processing[3].

### **4. Dataleon:**

API: Dataleon's Receipt Parser API offers high accuracy and real-time receipt management for data extraction, with customisable field selection and support for various receipt formats .

### **5. Google Cloud:**

API: Google Cloud's Receipt Parser API leverages machine learning to extract data from receipts with high accuracy, even handling handwritten receipts, and offers customizable solutions for specific data fields[4].

### **6. Azure:**

API: Azure's Receipt Parsing API, powered by the Form Recognizer receipt model, combines OCR and deep learning to extract information from various receipt formats and qualities, returning data in structured JSON format.

## **LICENSES / SUBSCRIPTIONS:**

### Document Intelligence

Azure Subscription: You need an active Azure subscription to use the Document Intelligence service.

Document Intelligence Instance: You need to create a Document Intelligence instance in the Azure portal to access the service.

API Keys and Endpoints: You need to obtain the API key and endpoint from your Document Intelligence resource to use the service.

### DocuClipper

Software License: You need a software license to use DocuClipper, which is a specialised web-based tool for converting invoices into structured data formats.

### Mindee

API Key: You need an API key to use Mindee's invoice OCR API, which extracts data from invoices with high accuracy and real-time processing.

### Dataleon

API Key: You need an API key to use Dataleon's Receipt Parser API, which offers high accuracy and real-time receipt management for data extraction.

### Google Cloud

API Key: You need an API key to use Google Cloud's Receipt Parser API, which leverages machine learning to extract data from receipts with high accuracy.

### Azure

API Key: You need an API key to use Azure's Receipt Parsing API, which combines OCR and deep learning to extract information from various receipt formats.

### Additional Requirements

Software Development Kits (SDKs): You can use various SDKs such as C# SDK, Python SDK, Java SDK, and JavaScript SDK to integrate the Document Intelligence invoice model into your applications.

Azure CLI: You can use the Azure CLI to create a Document Intelligence resource if you prefer to manage your resources through the command line.